Evolutionary analysis of rapidly evolving RNA viruses

Melissa J. Ward

Doctor of Philosophy
Institute for Evolutionary Biology
School of Biological Sciences
The University of Edinburgh
2012

Abstract

Recent advances in sequencing technology and computing power mean that we are in an unprecedented position to analyse large viral sequence datasets using state-of-the-art methods, with the aim of better understanding pathogen evolution and epidemiology. This thesis concerns the evolutionary analysis of rapidly evolving RNA viruses, with a focus on avian influenza and the use of Bayesian methodologies which account for uncertainty in the evolutionary process. As avian influenza viruses present an epidemiological and economic threat on a global scale, knowledge of how they are circulating and evolving is of substantial public health importance. In the first part of this thesis I consider avian influenza viruses of haemagglutinin (HA) subtype H7 which, along with H5, is the only subtype for which highly pathogenic influenza has been found. I conduct a comprehensive phylogenetic analysis of available H7 HA sequences to reveal global evolutionary relationships, which can help to target influenza surveillance in birds and facilitate the early detection of potential pandemic strains. I provide evidence for the continued distinction between American and Eurasian sequences, and suggest that the most likely route for the introduction of highly pathogenic H5N1 avian influenza to North America would be through the smuggling of caged birds.

I proceed to apply novel methods to better understand the evolution of avian influenza. Firstly, I use an extension of stochastic mutational mapping methods to estimate the d_N/d_S ratio of H7 HA on different neuraminidase (NA) subtype backgrounds. I find d_N/d_S to be higher on the N2 NA background than on N1, N3 or N7 NA backgrounds, due to differences in selective pressure. Secondly, I investigate reassortment, which generates novel influenza strains and precedes human influenza pandemics. The rate at which reassortment occurs has been difficult to assess, and I take a novel approach to quantifying reassortment across phylogenies using discrete trait mapping methods. I also use discrete trait mapping to investigate inter-subtype recombination in early HIV-1 in Kinshasa, the epicentre of the HIV-1 group M epidemic. In the final section of the thesis, I describe a method whereby epidemiological parameters may be inferred from viral sequence data isolated from different infected individuals in a population. To conclude, I discuss the findings of this thesis in the context of other evolutionary and epidemiological studies, suggest future directions for avian influenza research and highlight scenarios in which the methods described in this thesis might find further application.

Acknowledgements

This PhD was funded by the BBSRC. I am also grateful to ICHAIR for providing the opportunity for stimulating discussion with influenza researchers across Scotland.

I should like to thank my supervisors, Andy Leigh Brown and Sam Lycett, for providing me with the opportunity to undertake a PhD and for their incredible academic and pastoral support over the past few years. They have both put a considerable amount of their time into meetings and discussions which were not only invaluable for the progress of my project, but also in helping me to become so much more confident and independent in my research. Their patience and understanding has really been quite astounding.

I am hugely grateful to Jon Bollback for computational assistance at the start of my PhD, as well as for funding a very productive trip to visit his group in Vienna. I also thank Andrew Rambaut and Trevor Bedford for helpful technical discussions, and am indebted to Erik Volz for allowing me to become involved in his work. During my time in Edinburgh I have very much enjoyed discussions with past and present members of the Leigh Brown group: Fraser Lewis, Gareth Hughes, Lucy Weinert, Emma Hodcroft, Manon Ragonnet, Mojca Zelnikar and Lu Lu, as well as the Ashworth Virus Club.

I thank Jayna Raghwani, Laura Emery and Sema Nickbakhsh for friendships which have made the PhD process so much more fun. My office mates in Room 102, especially Jess Hedge and Matty Hartfield, have also provided good humour which has eased me through writing up. I am extremely grateful to my friends away from Edinburgh, especially Maddy, Sally, Bex, Lucy, Claire, Lara, Kirsty, Martin and Michael, for the perspective they are always happy to provide. I thank my family for their love, encouragement and belief in me. I must also pay tribute to staff at the Royal Infirmary of Edinburgh, without whom I would not have been able to undertake a PhD, and to Iola Wilson for helping me to keep going once I had started it.

Finally, I should like to thank Ed. He has put up with so much, and given so much, without ever tiring or complaining. With him, I feel that anything is possible.

Declaration

I declare that this thesis is my own composition and that the work described herein is my own, except where explicitly stated below. This work has not been submitted for any degree or professional qualification except as specified.

Melissa J. Ward

30 August, 2012

Chapter 4

The mutational mapping program used in this chapter was written by Dr. Jonathan Bollback as an adapted version of his SIMMAP software. Dr. Bollback devised the re-scalings for calculating d_N/d_S in consultation with me.

Chapter 5

The idea of using discrete trait mapping methods for investigating reassortment was put forward by Dr. Samantha Lycett and Professor Andrew Rambaut.

Chapter 6

Professor Andrew Rambaut provided advice on the HIV study presented in Chapter 6. Emma Hodcroft provided initial sequence alignments for Chapter 6.

Chapter 7

The mathematical results presented in this chapter were first derived by Dr. Erik Volz. I then derived the results from an initial outline of the results and reasoning, added detail and recommended amendments where necessary, and wrote explanatory material to help produce a manuscript of publication quality. This work was subsequently published as Volz *et al*. (2009) and also formed the basis of Volz *et al*. (2012). Dr. Trevor Bedford, Dr. Erik Volz and Nuno Faria were involved in discussions about the application of the method to HIV-1 group M in West Central Africa.

Contents

List of Figures

List of Tables

Chapter 1

Introduction to avian influenza

1 Introduction to avian influenza

In this chapter, an overview of the influenza virus is provided. In particular, the focus is on avian influenza viruses, their molecular and evolutionary genetics and the threat posed by their transmission to humans and animals.

1.1 Background to influenza

Influenza is a single stranded, negative-sense RNA virus belonging to the *Orthomyxoviridae* family (Smith *et al.* 1933; Palese and Shaw 2007). Type A influenza viruses are associated with much higher levels of diversity, infection and pathogenicity than type B and C influenza viruses and will be the subject of this investigation. In the twentieth century, the influenza A pandemics of 1918, 1957 and 1968 resulted in significant human mortality. Introduction of genetic material from avian influenza viruses led to the 1957 and 1968 pandemics, and has also been implicated in the origin of the 1918 pandemic virus. The 2009 H1N1 'swine 'flu' pandemic highlighted the constant threat that human populations are under from the emerging influenza strains. The large reservoir of avian influenza viruses currently circulating, coupled with the previous periodic emergence of pandemic strains of avian origin, has led to concern that an avian influenza virus could give rise to another pandemic (Webby and Webster 2003).

Given their role in past pandemics, it is important to understand how avian influenza viruses are circulating and evolving in birds, in order to target surveillance studies and pre-empt the emergence of pandemic strains. It is unclear which geographical region a new pandemic influenza virus will come from, and what subtype it will be. Indeed, the 2009 H1N1 swine-origin pandemic strain (H1N1-2009) was not identified as a threat until it had caused widespread human infection which could not be contained (Fraser *et al.* 2009; Smith *et al.* 2009). In addition to the epidemiological and economic consequences of human infection, influenza viruses are associated with profound economic losses to the poultry industry as a single avian influenza outbreak can lead to the slaughter of several million birds. Understanding the genetic factors underlying the emergence of avian influenza viruses which cause particularly severe infection in humans, other mammals and birds could facilitate the early identification of strains with the potential to cause widespread or severe disease and influence strategies for their control.

1.2 Structure and function of influenza viruses

The genome of the influenza A virus is organised into eight RNA segments (Palese 1977), numbered from 1 to 8, from which a total of eleven proteins are transcribed (Chen *et al.* 2001). An influenza virion must contain a copy of each of the eight RNA segments to be fully infectious (Palese and Shaw 2007). The names and functions of the avian influenza virus proteins, as well as the number of the RNA segment encoding the protein, are listed in Table 1.1. Figure 1.1 is a cross-sectional representation of an influenza virion. The glycoproteins haemagglutinin (HA) and neuraminidase (NA) may be observed protruding in spikes at the viral surface. The structural matrix protein layer is found below the surface layer of the virion. At the core of the virion lies the helical nucleoprotein-RNA-polymerase complex, which contains copies of the eight RNA segments, the polymerase proteins required for viral RNA synthesis, the nucleoprotein, which has structural and regulatory roles, and the non-structural proteins required for viral replication (Webster *et al.* 1992).

Protein Name	Number of RNA segment encoding protein	Function
Haemagglutinin (HA)	4	Attachment to host cell
Neuraminidase (NA)	6	Release of viral progeny from host cell
Nonstructural NS1 Nonstructural NS2	8	Viral replication; NS1inhibits host innate immune response
PB1 Polymerase	2	Viral RNA synthesis
PB1(F2) Polymerase	2	
PB2 Polymerase	1	
PA Polymerase	3	
Nucleoprotein (NP)	5	Structural and regulatory
M1 Matrix M2 Matrix	7	Virus structure

Table 1.1
Avian influenza A virus proteins. 11 proteins are coded for by a total of 8 RNA segments (numbered 1-8, in order of length, with 1 being the largest) in the influenza A virus genome. The main function of the protein, and the number of the RNA segment encoding it, are listed (information from Webster *et al.* 1992; PB1-F2 from Chen *et al*. 2001)

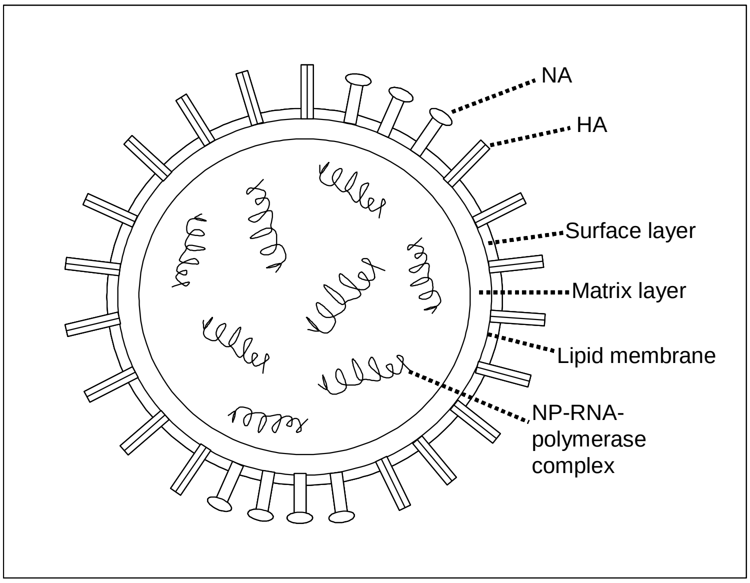
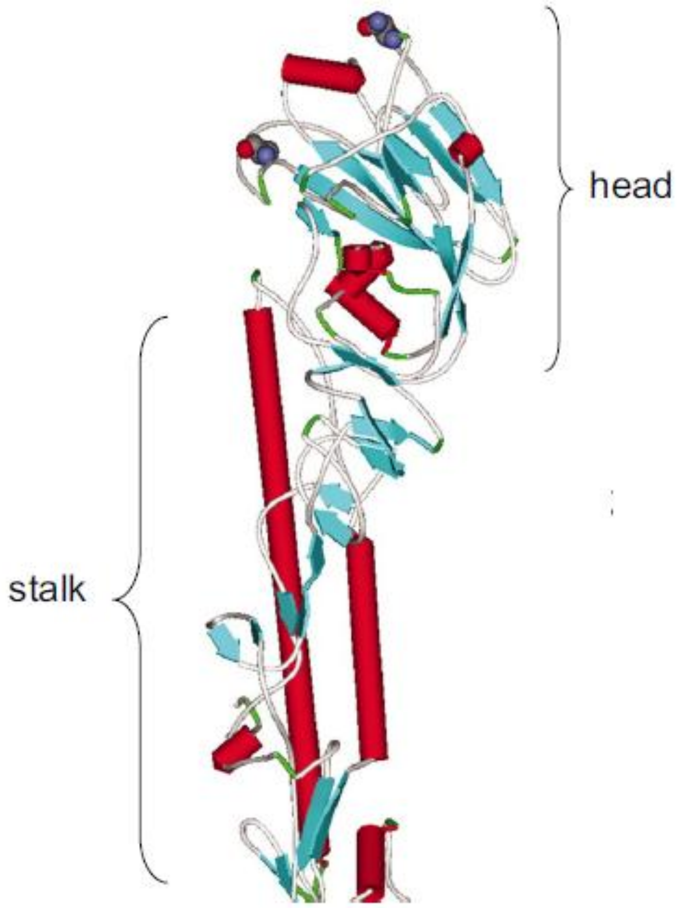


Figure 1.1
Cross-sectional representation of an influenza virion. The surface glycoproteins HA and NA are responsible for entry into and release from host cells respectively, and protrude in spikes at the virion surface. Beneath the surface layer lies the matrix layer, which gives structure to the virion. The helical nucleoprotein-RNA-polymerase complex, which contains copies of the viral RNA, can be seen at the core of the virion.

HA binds to sialic acid receptors in the host and mediates entry into the host cell; for review see Wiley and Skehel (1987). HA is the major antigenic influenza protein (Skehel and Wiley 2000) and current influenza vaccines primarily target the HA protein to elicit a protective antibody response (Eckert and Kay 2010). Post-translational cleavage of the HA precursor protein HA0 into subunits HA1 and HA2 is necessary for viral infectivity; for review see Steinhauer (1999). The globular head of the HA protein (Figure 1.2) is encoded by the HA1 subunit and is the main antigenic domain of the influenza HA. Five antigenic epitopes have been identified for H1 (Gerhard *et al.* 1981) and H3 HA (Webster and Laver 1980; Wiley *et al.* 1981), denoted A to E for H3 and Ca1, Ca2, Cb, Sa and Sb for H1 (see below for information on the biological meaning of HA subtypes H1 and H3). The amino acid sites corresponding to these epitopes have been recorded. The receptor binding site, which binds to sialic acid receptors in host cells, is also encoded by the HA1 sub-unit. Most of the molecular variation in influenza HA is found in the HA1 region, whilst HA2, which along with short regions of the HA1 encodes the stalk of the protein, has been shown to be relatively conserved (Krystal *et al.* 1982).

NA also binds to sialic acid receptors, facilitating the release of progeny virions from host cells (Seto and Rott 1966) and preventing aggregation below the cell surface (Palese *et al.* 1974). NA has a major antigenic role for the influenza virus, second in importance only to the HA, and at least two antigenic sites have been identified for NA (Colman *et al.* 1983; Air *et al.* 1985). Neuraminidase-inhibiting drugs (NAIs), which block the enzyme activity of NA and curb the spread of virus in the host by preventing the release of progeny virions from infected cells, are used in the treatment of influenza in humans. NAIs include the orally administered oseltamivir (Tamiflu), which was widely used during the 2009 H1N1 pandemic, and zanamivir, which is inhaled orally or nasally. The other major class of anti-influenza agents is the adamantanes (rimantadine and amantadine), which interfere with the M2 matrix protein ion channel (Wang *et al.* 1993; Pinto and Lamb 1995). Widespread amantadine resistance has arisen amongst seasonal influenza strains (Deyde *et al.* 2007) and all pandemic H1N1-2009 viruses were amantadine resistant (Dawood *et al.* 2009). Although mutations exist which are known to confer resistance to NAIs (note, for example, the recent global spread of oseltamivir–resistant seasonal H1N1 amongst humans (Moscona 2009)), NAIs remain the most effective treatment for influenza infection.



Hoffmann E et al. PNAS 2005;102:12915-12920

Figure 1.2
Three-dimensional structure of an influenza haemagglutinin (HA) molecule. The globular head is encoded by the HA1 subunit and contains the main epitopes and the receptor binding site of the protein. The stalk region is largely encoded by the more conserved HA2 subunit and mediates entry of the virus into host cells.

The HA and NA proteins are grouped into subtypes which are not serologically cross-reactive (WHO 1980). There are currently 16 known HA subtypes (H1 – H16) and 9 known subtypes of NA (N1 – N9) circulating in avian populations (Webster *et al.* 1992; Fouchier *et al.* 2005). Protein coding sequences for the HA1 sub-unit have been shown to vary by at least 30% between different HA subtypes (Webster *et al.* 1992). Phylogenetically, the HA subtypes form two distinct clades: group 1, consisting of subtypes H1, H2, H5, H6, H8, H9, H11, H12, H13 and H16 and group 2, comprising subtypes H3, H4, H7, H10, H14 and H15 (Air 1981; Chen and Holmes 2006). Similar observations have been made for NA, with the N1, N4, N5 and N8 subtypes forming a distinct phylogenetic cluster (denoted group 1) and N2, N3, N6, N7 and N9 clustering together (denoted group 2) (Fouchier *et al.* 2005). X-ray crystallography has revealed that the two groups of NA subtypes are structurally distinct (Russell *et al.* 2006).

The failure of neutralising antibodies to cross-react with subtypes from different clades, whilst providing some protection against subtypes from the same group, provides support for serological and structural differences between the two HA subtype groups (Okuno *et al.* 1993; Kashyap *et al.* 2008; Ekiert *et al.* 2009; Sui *et al.* 2009). Further division of HA subtypes into four clades, each represented by one of subtypes H3, H5, H7 or H9, has been supported phylogenetically (Nobusawa *et al.* 1991) and evidence has been provided for structural differences between the HA subtypes belonging to the four clades (Ha *et al.* 2002; Russell *et al.* 2004).

1.3 Influenza virus nomenclature

Influenza A viruses are classified by their HA and NA subtypes, so that a virus of HA subtype H5 and NA subtype N1 would be referred to as being of the H5N1 serotype. The standard nomenclature for influenza viruses is to include the type of virus (A, B or C), the host of origin (except for human influenza viruses, where this is omitted), the geographic site of sampling, the strain number and the year of sampling, followed by the subtype in parentheses (Wright *et al.* 2007). Thus, A/goose/Guangdong/1/96(H5N1) would indicate the first strain of a type A virus of H5N1 subtype to have been isolated from a goose in the Guangdong province of China in 1996.

1.4 Birds as the natural hosts of the influenza A virus

Wild aquatic birds of orders *Anseriformes* (e.g. ducks, geese and swans) and *Charadriiformes* (e.g. gulls, terns, surfbirds and sandpipers) have been established as the natural reservoir for type A influenza viruses (Wright *et al.* 2007). All known HA and NA subtypes have been found to exist in waterfowl (Webster *et al.* 2007), although the most prevalent subtypes of influenza virus isolated from gulls and shorebirds differ from those which predominate in ducks (Webster *et al.* 1992; Krauss *et al.* 2004). The influenza virus is enteric in avian species and is shed in high quantities in the faeces of avian hosts (Webster *et al.* 1978). Faecal shedding in lakes is a major mechanism for the transmission of influenza viruses amongst wild waterfowl (Webster *et al.* 1992). Although influenza infection is largely asymptomatic in ducks, sporadic occurrences of pathogenic infections have been reported in these hosts (Sturm-Ramirez *et al.* 2004), particularly associated with recent H5N1 viruses (Chen *et al.* 2005; Liu *et al.* 2005) (see Section 1.6). Avian influenza viruses which are virulent in *Charadriiformes* have occasionally been reported (Becker 1966).

Annual peaks in influenza virus infections amongst wild waterfowl have been shown to occur after the breeding season and are accounted for by high levels of infection amongst juvenile birds, which are immunologically naïve (Hinshaw *et al.* 1980). It is not known exactly how influenza viruses persist in their natural host populations at times of the year when infection levels are very low. Two main mechanisms have been put forward for the perpetuation of influenza viruses in wild migratory birds (Stallknecht *et al.* 1990) and evidence for both exists. Firstly, influenza viruses may become frozen in the waters at the breeding grounds of avian species and be able to re-infect when they thaw the next year (Ito *et al.* 1995). Secondly, influenza viruses could be maintained at low levels in flocks of migratory birds over the winter (Krauss *et al.* 2004). In addition, it is possible that influenza viruses could circulate continuously amongst *Anseriformes* and *Charadriiformes* in subtropical and tropical regions (Webster *et al.* 1992); indeed, a tropical epicentre has also been suggested for human influenza viruses (Rambaut *et al.* 2008).

Influenza viruses from wild birds separate into phylogenetically distinct clades of Eurasian and American viruses (Olsen *et al.* 2006) and the opportunity for transatlantic genetic exchange along migratory flyways is thought to be limited (Figure 1.3). Indeed, despite the prevalence of the highly pathogenic H5N1 virus amongst poultry in Asia, Africa and Europe, this strain has not been detected in North or South America. Although transfer between viruses in the Eurasian and American clades has been documented for certain segments (Schafer *et al.* 1993; Makarova *et al.* 1999; Widjaja *et al.* 2004), phylogenetic evidence suggests that such events occur at a low frequency (Olsen *et al.* 2006; Webster *et al.* 2007).

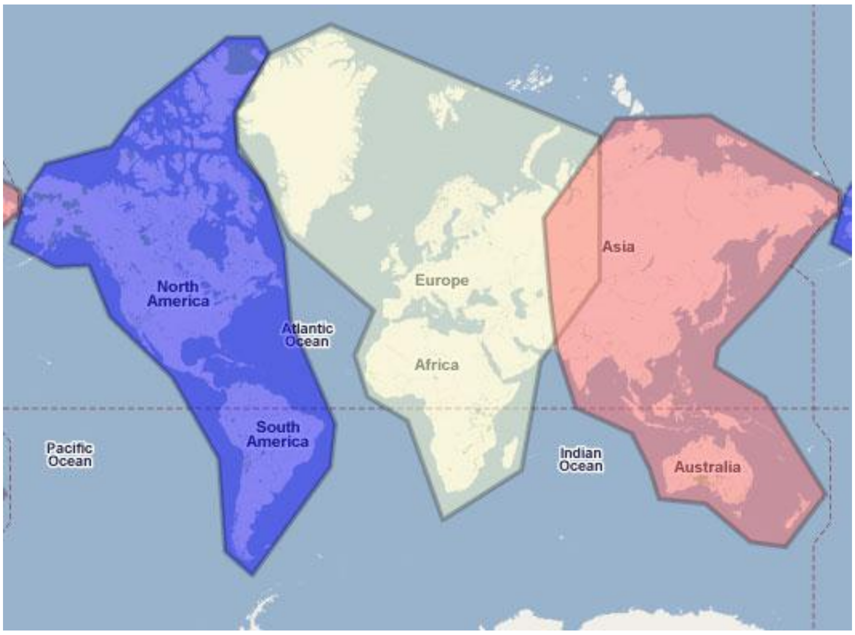


Figure 1.3
Migratory flyways of wild birds. Three major global flyways of wild migratory birds have been identified: the Americas flyway (shown in blue), the African-Eurasian flyway (yellow) and the East Asian-Australasian flyway (red). Image obtained, with permission, from http://www.birdlife.org. Examination of these flyways suggests the possibility for exchange of avian influenza viruses between Europe and Asia, whilst exchange of virus between the Americas and Eurasia is thought to be more limited.

1.5 Influenza in poultry

Influenza viruses have been found to infect various types of domestic poultry, such as chickens, turkeys, ducks and geese. Outbreaks of avian influenza in poultry, whether of the low pathogenic or highly pathogenic form (see Section 1.6), can result in economic loss and disruption to the poultry industry through attempts to contain or eradicate the virus by culling or vaccination, as well as the implementation of local and international trade restrictions (Davison *et al.* 1999; Pearson 2003). Most influenza viruses infecting domestic poultry are thought to have been transmitted directly from wild birds, although evidence exists for the transmission of influenza viruses from pigs to turkeys (Hinshaw *et al.* 1983; Choi *et al.* 2004).

Influenza viruses have been isolated sporadically from passerine birds, and starlings and sparrows have been implicated in the spread of H7N7 amongst poultry in Australia (Nestorowicz *et al.* 1987). Evidence for the repeated introduction of avian influenza viruses from wild birds into domestic poultry populations has been provided by phylogenetic analyses, which have found that sequences isolated from wild birds and domestic outbreaks do not form distinct clades. For example, Banks *et al*. showed that early sequences from the 1999 H7N1 poultry outbreak in Italy clustered most closely with a wild bird sequence from Taiwan, rather than with contemporaneous Eurasian poultry isolates (Banks *et al.* 2001).

Avian influenza viruses may be transmitted between poultry flocks as a result of poor biosecurity measures during the movement of poultry between live bird markets, farms and slaughterhouses (McQuiston *et al.* 2005). Surveillance has indicated that influenza viruses are endemic in certain poultry populations, such as in the live bird markets of South-East Asia (Liu *et al.* 2003) and North America (Senne *et al.* 2003b; Suarez *et al.* 2003). Live bird markets, in which multiple influenza subtypes circulate and where the virus can persist and replicate in poultry for extended periods of time, are thought to be an ideal environment for the emergence of virulent avian influenza viruses (Senne *et al.*

2003a; Webster 2004; Cardona *et al.* 2009). Examples of avian influenza outbreaks in poultry for which the live bird markets have been implicated as the source include the 1983-1984 H5N2 outbreak in North-East America (Suarez and Senne 2000) and multiple outbreaks of H7N2 in the United States during the 1990s (Akey 2003; Davison *et al.* 2003; Dunn *et al.* 2003).

1.6 Highly pathogenic avian influenza (HPAI)

Avian influenza viruses are classified as being of high pathogenicity (HPAI or HP) or low pathogenicity (LPAI or LP) according to their effects in chickens, with HPAI viruses being highly virulent and associated with considerably more severe symptoms than LPAI. The following are the recognised guidelines regarding the conditions under which avian influenza virus should be classified as HPAI (Alexander 2000):

(a) Any influenza virus that is lethal for six, seven or eight of eight 4-6 week old susceptible chickens within 10 days following intravenous inoculation with 0.2ml of a 1/10 dilution of a bacteria-free, infective allantoic fluid

(b) The following additional test is required if the isolate kills from one to five chickens, but is not of the H5 or H7 subtype: growth of the virus in cell culture with cytopathic effect or plaque formation in the absence of trypsin. If no growth is observed, the isolate is not considered to be a HPAI isolate

(c) For all H5 and H7 viruses of low pathogenicity and for other influenza viruses, if growth is observed in cell culture without trypsin, the amino acid sequence of the connecting peptide of the haemagglutinin must be determined. If the sequence is similar to that observed for other HPAI isolates, the isolate being tested will be considered to be highly pathogenic.

Early twentieth century isolates of HPAI were termed ‘fowl plague virus’ (FPV) until they were identified as influenza viruses by Schafer (1955). Avian influenza viruses not designated HPAI are classified as LPAI and are typically associated with symptoms such as mild respiratory disease and a decrease in egg production rather than mortality.

Viruses of HA subtypes H5 and H7 have been associated with all major recorded outbreaks of HPAI, and the overwhelming majority of recorded outbreaks of HPAI have occurred in poultry (Wright *et al.* 2007). Table 1.2 lists all documented outbreaks of HPAI as of 31st December, 2010 and gives the subtype, location and year of each outbreak; this illustrates the continuing potential for virulent avian influenza viruses of different subtypes to infect poultry in diverse geographic regions. Reports of H5N1 HPAI since 2003 are too numerous to include in the table, but comprise over 5,000 outbreaks in Asia, Africa and Europe (WAHID 2011). Outbreaks of HPAI viruses lead to the deaths of millions of birds every year as a result of infection and control measures. In addition, HPAI viruses present a global epidemiological threat should they become established in the human population (Peiris *et al.* 2007), although this has not happened to date, with known human pandemics restricted to the H1, H2 and H3 HA subtypes (Webster *et al.* 1992; WHO 2010).

Year of start of outbreak	Subtype	Location	Additional details
1959	H5N1	Scotland	
1961	H5N3	South Africa	
1963	H7N3	England	
1966	H5N9	Ontario	
1976	H7N7	Victoria	
1979	H7N7	Germany	
1979	H7N7	England	
1983	H5N2	Pennsylvania	
1983	H5N8	Ireland	
1985	H7N7	Australia (Victoria)	
1991	H5N1	England	
1992	H7N3	Australia (Victoria)	
1995	H5N2	Puebla	
1995	H7N3	Australia (Queensland)	
1995	H7N3	Pakistan	
1997	H5N1	Hong Kong	
1997	H7N4	New South Wales	
1997	H5N2	Italy	
1999	H7N1	Italy	
2002	H7N3	Chile	
2002	H5N1	Hong Kong	Wild birds
2003	H7N7	Netherlands	
2004	H7N3	Canada	
2004	H5N2	South Africa	
2004	H5N2	Chinese Taipei	
2005	H7N7	Korea (Dem. People's Rep.)	
2005	H5N2	Zimbabwe	
2006	H5N2	South Africa	
2007	H7N3	Canada	
2008	H7N7	United Kingdom	
2009	H7N7	Spain	

Table 1.2
Recorded outbreaks of highly pathogenic avian influenza. Data are adapted from Wright *et al*. (2007) for outbreaks prior to 2004, and the World Animal Health Information Database Interface (WAHID) (accessed 19th January, 2011) for data from 2004 onwards. H5N1 outbreaks since 2003 are too numerous to record in the table and the strain is now endemic in regions of South East Asia and North Africa. Over 5,000 outbreaks of HPAI H5N1 were reported to the WAHID between 2003 and February 2001.

LPAI viruses can mutate into HPAI viruses in poultry, as was documented for outbreaks of H5N2 in Pennsylvania in 1983 (Bean *et al.* 1985; Suarez *et al.* 2004) and H7N1 in Italy in 1999 (Banks *et al.* 2001). Molecular changes resulting in a shift in virulence have also been documented in laboratory experiments involving serial passage in chicken embryo cells (Horimoto and Kawaoka 1995). Most HPAI viruses contain multiple basic amino acids (histidine (H), lysine (K) or arginine (R)) at the HA cleavage site (Perdue *et al.* 1997; Steinhauer 1999) and a minimum motif of B-X-X-B-R, where B denotes arginine or lysine and X denotes a non-basic amino acid, has been proposed (Wood *et al.* 1993; Senne *et al.* 1996). It is thought that these molecular changes allow the HA0 precursor protein to be cleaved into HA1 and HA2 in tissues throughout the body, rather than just in the respiratory and gastrointestinal tracts, leading to systemic infection and the neurological symptoms associated with HPAI infection (Horimoto and Kawaoka 2001).

Reverse genetics studies have provided direct evidence for the link between HA cleavability and virulence of avian influenza viruses (Horimoto and Kawaoka 1994). Recent evidence suggests that mutations in the regions adjacent to the HA cleavage site also play a part in determining the pathogenicity of avian influenza viruses (Gohrbandt *et al.* 2011). Whilst ubiquitous HA cleavability is a necessary condition, pathogenicity of influenza viruses in chickens is thought to be a polygenic trait (Rott *et al.* 1976), determined at least by the genes involved in viral RNA synthesis in addition to the HA; for example, see Rott (Rott 1980). Experimental evidence suggests that virulence of avian influenza viruses in mammals is also determined by a number of genes (see Section 1.8).

HPAI influenza viruses of the H5N1 subtype have come to prominence in the scientific community and the media as a result of their sporadic transmission to humans with high mortality rates (WHO 2005). During an H5N1 outbreak in Hong Kong in 1997, which was controlled by culling all poultry on the island, the first fatal direct transmission of the virus from chicken to human was recorded (Claas *et al.* 1998; Subbarao *et al.* 1998). A total of 306 fatalities out of 518 laboratory-confirmed human infections with HPAI H5N1 have been reported to the World Health Organisation (WHO) as of 20th January 2011 (WHO 2011a). The H5N1 HPAI outbreak at Qinghai Lake in China in 2005 resulted in the deaths of over 6000 migratory waterfowl (Chen *et al.* 2005) and demonstrated that some highly pathogenic influenza viruses could establish outbreaks in populations of wild birds (Liu *et al.* 2005). The virulence of HPAI H5N1 has been found to vary in aquatic waterfowl and it has been suggested that ducks in which infection is associated with less severe symptoms may propagate the HPAI H5N1 virus (Hulse-Post *et al.* 2005). Experimental studies have shown that susceptibility to HPAI H5N1, clinical disease and viral shedding may differ between species of wild birds (Brown *et al.* 2006; Brown *et al.* 2008). Surveillance studies have indicated that highly virulent H5N1 viruses are endemic in poultry in southern China and will be difficult to eradicate (Li *et al.* 2004). Endemic HPAI H5N1 infection has also been declared in Egypt, Bangladesh, Indonesia and Viet Nam (WHO 2011b).

Despite the endemic nature of HPAI H5N1 in many regions, and the high levels of mortality associated with human infection, this subtype must not be considered the only avian influenza strain with pandemic potential. Indeed, the next human pandemic virus of avian origin need not be highly pathogenic in birds and such strains are more likely to go unnoticed by surveillance studies (Peiris *et al.* 2007). The avian influenza viruses which contributed segments to the 1957 and 1968 pandemic viruses were not highly pathogenic in poultry. Furthermore, the H1N1-2009 virus was only associated with moderate symptoms in humans, yet possessed the ability to efficiently transmit between humans and cause a global influenza outbreak. Avian H9N2 viruses have been identified as pandemic candidates due to their widespread presence in poultry in Asia (Matrosovich *et al.* 2001), their gradual spread westwards (Alexander 2003) and the continued reassortment in the matrix and NP genes of certain strains (Choi *et al.* 2005). Also, H9N2 viruses isolated from poultry in Hong Kong have been found to exhibit human-like receptor-binding preferences whilst maintaining the ability to infect other avian species (Matrosovich *et al.* 2001).

1.7 Influenza viruses in non-avian species

Influenza viruses infect a variety of non-avian species, including humans, swine, horses and other mammals. Seasonal epidemics of H1N1 and H3N2 influenza in humans typically take place during the winter months and result in mortality amongst the elderly and infirm, coupled with economic loss due to workforce absence. In addition to seasonal influenza epidemics, human populations are susceptible to pandemic influenza viruses which arise sporadically and spread to cause high levels of infection globally. In addition to the three pandemics of the twentieth century, and the 2009 pandemic, influenza pandemics were described in the nineteenth century with the latest being in 1890 (Taubenberger and Morens 2006).

Viruses of subtype H1N1 were responsible for the 1918 influenza pandemic (Reid *et al.* 1999), which led to an estimated 50 million deaths worldwide (Johnson and Mueller 2002). Whilst avian viruses have been suggested as the causative agents of the 1918 H1N1 pandemic strain, the origin and genetic basis of this virus are yet to be fully understood (Reid and Taubenberger 2003). The H2N2 virus responsible for the 1957 influenza pandemic is thought to contain HA and NA genes of avian origin (Scholtissek *et al.* 1978) on a genetic background from the H1N1 strain. By 1968, the H2N2 viruses had been replaced by H3N2 viruses with an avian H3 HA segment (Scholtissek *et al.* 1978; Nakajima *et al.* 1982). The PB1 segment of the 1957 pandemic strain, and that which replaced it in 1968, were both found to be of avian origin (Kawaoka *et al.* 1989). These findings highlight the potential of avian influenza viruses to contribute genetic material to a future pandemic strain.

Influenza viruses of subtypes H1N1 and H3N2 have become established in swine and phylogenetic studies have shown that they form a sister group to human influenza viruses (Webster *et al.* 1992). Prior to the 2009 H1N1 pandemic, porcine influenza viruses were known to be transmissible to humans and had been shown to occasionally lead to respiratory symptoms in humans (Dasco *et al.* 1984).

Phylogenetic analysis indicated that the H1N1-2009 pandemic strain possessed genetic components from multiple swine lineages; the HA, NP, NS and polymerase genes were contributed by a 'triple reassortment' lineage which was circulating in swine and was itself derived from human, avian and classical swine viruses (Smith *et al.* 2009). The progenitor H1N1-2009 virus was not detected by swine surveillance programmes and it is thought that the initial movement of the virus from pigs to humans could have taken place several months before the outbreak was reported (Smith *et al.* 2009).

Although associated with less severe symptoms than the 1918 H1N1 pandemic strain, the H1N1-2009 virus caused approximately 18,000 recorded human deaths within a year of the pandemic being announced (WHO 2010). Children and young adults accounted for the highest proportion of symptomatic infections, whilst those in the over 55 age group displayed remarkably low levels of symptomatic infection and the highest levels of antibody response, suggesting previously acquired immunity (Hancock *et al.* 2009). The pandemic H1N1-2009 virus was the most widely circulating strain during the seasonal peak in influenza infections in the late part of 2010; as with the first wave of the pandemic in 2009, young adults and children were most likely to be affected (Ellis *et al.* 2011). Speculation remains as to whether the continued circulation of H1N1-2009 will be prevented by high levels of pre-existing and vaccine-induced immunity and what part it will play in future seasonal epidemics (Morens *et al.* 2010), although surveillance has revealed that it is co-circulating with H3N2 in the 2012 season, with H3N2 variants containing a matrix segment from the H1N1 pandemic virus having also been detected (CDC 2012).

H3N8 and H7N7 influenza viruses are thought to have persisted in equine populations for several hundred years (Wright *et al.* 2007). Equine influenza viruses continue to be isolated to this day, with a recent example being the finding of H7N7 amongst polo horses in Nigeria (Olusa *et al.* 2010). Phylogenetic analyses suggest that modern-day equine populations are at risk from novel influenza viruses entering the population from the avian reservoir. For example, a distinct virus of avian origin was responsible for the Chinese equine influenza outbreak in 1989 (Guo *et al.* 1992).

Other documented mammalian influenza infections with non-H5N1 influenza in the natural world include that of seals along the west coast of North America with H7N7 viruses in 1979-1980 (Geraci *et al.* 1982) and, more recently, H3N8 infection in North American dogs (Crawford *et al.* 2005).

1.8 Virulence of HPAI in mammals

HPAI isolates are, by definition, virulent in chickens (see Section 1.6). However, their effects have been found to vary in mammals, with some strains causing only mild symptoms in these hosts (Katz *et al.* 2000; Govorkova *et al.* 2005). Experimental studies have investigated the virulence of H5N1 HPAI in mammals, particularly as this subtype has been associated with high mortality in humans. Evidence suggests that virulence of influenza viruses in mammals is a polygenic trait, involving at least the non-structural (Zamarin *et al.* 2006; Conenello *et al.* 2007) and polymerase proteins (Hale *et al.* 2008; Jackson *et al.* 2008) as well as the HA. Specific mutations have been found to contribute to virulence, for example a mutation to lysine at site 627 in PB2 has been shown to confer virulence in reverse genetics studies on mice (Hatta *et al.* 2001, 2007). However, such studies are resource-intensive and have typically been conducted on a small scale. The genetic basis of virulence of HPAI H5N1 and HPAI viruses of other subtypes is thus yet to be fully explained.

In order to investigate the molecular determinants of H5N1 HPAI virulence in mammals, Lycett *et al*. (2009) (see Chapter 11) performed a meta-analysis of data from studies where virulence in mammals had been determined experimentally (predominantly in ferrets and mice). Bayesian Graphical Models (BGMs) were used to investigate associations between amino acid site mutations and whether the virus was virulent in mammals. BGMs also highlight dependencies between mutations at different amino acid sites. This permits the identification of genetic constellations putatively related to virulence. BGMs represent a set of variables as nodes (in the Lycett *et al*. (2009) study, nodes represented amino acid sites, or the binary trait variable of ‘virulent’ or ‘non-virulent’) and dependencies between variables as directed edges in a directed acyclic graph (Pearl 1986). The aim of a BGM analysis is to distinguish true dependencies between variables from observed ‘correlations’ resulting from dependence on an intermediate node.

The directed acyclic graph obtained by Lycett *et al*. (2009) in analysing the genetic determinants of HPAI H5N1 virulence in mammals is shown in Figure 1.4. Three nodes were identified as being directly associated with virulence: those representing amino acid sites PB1-317/PB2-355, NS1-92/NS1-228 and HA-102/NS1-195 (note that for each of these three variables there are two sites for which the association is identified – these cannot be distinguished as the same pattern of mutations across the sequences was observed for these pairs of sites). Amino acid changes at site 92 in the NS1 protein have previously been implicated as a determinant of H5N1 virulence in mammals. Specifically, the mutation D92E in NS1 is thought to enable the virus to resist the host immune response (Seo *et al.* 2002; Li and Wang 2007). In addition, mutations at PB1-317 and PB2-355 had previously been identified experimentally as being correlated with H5N1 virulence in mice (Katz *et al.* 2000).

A further five amino acid sites (three in HA and two in PB2), as well as the presence/absence of a deletion in the stalk region of NA, were identified by Lycett *et al*. (2009) as being associated with one or more of the three nodes directly linked to virulence. One such site was PB2-627; the presence of a lysine at this site is thought to increase the ability of the virus to replicate in mammalian cells (Subbarao *et al.* 1993). The finding of strong associations between mutations in the HA and a deletion in the NA stalk region is not surprising, given existing evidence for concomitant changes in these proteins and their interacting functional roles (see Chapter 4, Section 4.3 for further discussion). Association between an NA stalk deletion and changes conferring additional glycosylation sites in the HA are also thought to play a role in the adaptation to domestic poultry of influenza viruses from wild aquatic birds (Matrosovich *et al.* 1999).

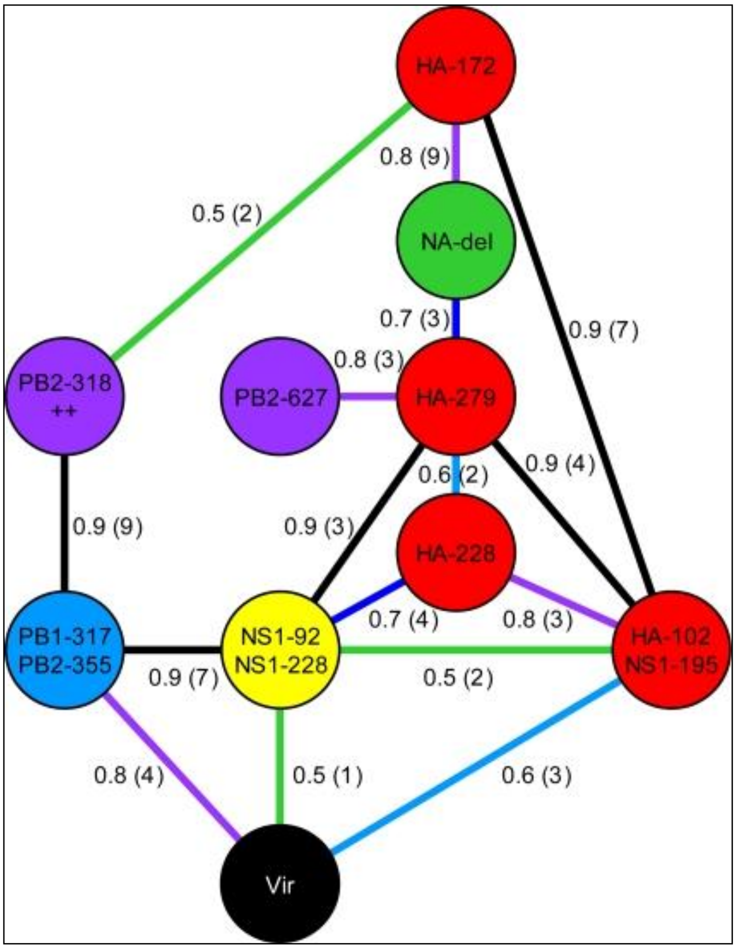


Figure 1.4
Bayesian graphical model (BGM) for the virulence of highly pathogenic H5N1 avian influenza viruses in mammals (Lycett *et al.* 2009). The nodes represent amino acid sites or the phenotype of 'virulent or avirulent in mammals' (node denoted *Vir*). Links indicate a probabilistic dependence between nodes and are labelled by their probabilities, inferred using 10-fold cross-validation. Three sites were found to be directly associated with virulence in mammals: PB1-317/PB2-355, NS1-92/NS1-228 and HA-102/NS1-19 (where more than one site is listed for each node, the same pattern of variation from the consensus sequences was observed for these pairs of sites).

1.9 Transmission of influenza viruses between host species

Despite the vast genetic diversity of avian viruses, only a few influenza subtypes have become established in human populations over the last century. The mechanisms by which influenza viruses infect and are maintained in different host species has been the subject of continued research. Three interacting processes have been identified as potential barriers between the infection of a novel (recipient) species with an influenza virus from a (donor) host in which the virus is already endemic (Kuiken *et al.* 2006): (i) interactions between the donor and recipient host species; (ii) virus-host interactions in the recipient species and (iii) host-host interactions within the recipient species. It is thus possible to identify high-risk interactions for the emergence of influenza viruses with the potential to cause human outbreaks. For example, the pattern of 'backyard' poultry-rearing or close contact between humans and poultry in live bird markets may facilitate the transmission of avian influenza viruses to humans. Furthermore, regular oral contact between humans and birds is known to occur in certain communities, for example during poultry feeding in Egypt or the rearing of fighting cocks in Thailand (ISID 2004); these are clearly high-risk interactions for influenza virus transmission from birds to humans.

Research into virus-host interactions has indicated that differences in receptor binding preference of the HA protein provide a transmission barrier between host species. Avian influenza viruses preferentially bind to α-2,3 linked sialic acid receptors in the host, whereas human influenza viruses preferentially bind to α-2,6 linked sialic acid receptors (Rogers and Paulson 1983; Matrosovich *et al.* 1997). Preference for α-2,6 linked receptors is thought to be important for efficient transmission of influenza between humans, a thesis which is supported by the findings that the 1957 and 1968 pandemic viruses share this preference, despite possessing HA segments of avian origin (Matrosovich *et al.* 2000). However, human-to-human transmission of H5N1 influenza remains limited, despite these viruses showing an affinity for α-2,6 linked receptors (Yamada *et al.* 2006). Thus, additional viral changes may be required for efficient transmission between humans.

For example, an amino acid substitution from glutamic acid to lysine at PB2 site 627 has been suggested to increase viral excretion via coughing and sneezing, thus helping the virus to spread from human to human (Hatta *et al.* 2007).

For many years before the H1N1-2009 pandemic, swine were highlighted as potential vessels in which a pandemic virus of avian origin could emerge in a form which is transmissible between humans (Scholtissek *et al.* 1985). This is due to the finding that porcine epithelial cells contain both α-2,3 and α-2,6 linked receptors (Ito *et al.* 1998). It has been shown that pigs can become infected with avian influenza viruses and that even avian viruses which are unable to replicate in pigs could contribute gene segments to viable reassortant viruses (Kida *et al.* 1994). The threat of the pig as an intermediate vessel for the emergence a future pandemic virus with an avian component is heightened by the finding that H3N2 viruses of human origin are co-circulating with avian influenza viruses in pigs in southern China (Peiris *et al.* 2001). The reported direct transmission of equine H3N8 viruses to greyhounds via feeding on horsemeat has even led to speculation that domestic dogs might provide a route for influenza viruses into the human population (Crawford *et al.* 2005).

1.10 Evolution of avian influenza viruses

Three main mechanisms exist by which influenza viruses may evolve: reassortment, intragenic recombination and the accumulation of single amino acid mutations. The segmented nature of the influenza genome means that, when a host is co-infected with two or more strains of the influenza virus, novel combinations of gene segments may be created during viral replication. Progeny viruses may be produced whose sets of eight RNA segments are derived from more than one ‘parent’ virus, resulting in a virus that is different than either of the original infecting strains. This process is referred to as reassortment. The evolution of influenza viruses by reassortment in the HA or NA segments is termed antigenic shift, whereas evolution via amino acid mutations is termed antigenic drift, to reflect the gradual nature of the latter compared to the former (Webster *et al.* 1982). Note that antigenic drift should not be confused with random genetic drift, as antigenic drift is driven by selective pressure from the host's immune system.

Reassortment in the antigenic HA and NA segments produces novel influenza subtypes, to which the majority of a host population will be naïve. High levels of reassortment in the HA and NA segments have been documented amongst avian influenza viruses isolated from wild birds (Dugan *et al.* 2008). Acquisition of a novel HA segment could confer an advantage to the influenza virus by enabling it to evade the host immune response (Palese and Shaw 2007), as could a novel NA segment (Gong *et al.* 2007). Reassortant viruses were responsible for the influenza pandemics of 1957 and 1968, with both pandemic viruses acquiring HA and PB1 segments originating in wild waterfowl (Scholtissek *et al.* 1978; Kawaoka *et al.* 1989). Reassortment amongst seasonal H1N1 influenza viruses has been implicated in the unusually severe epidemics of this subtype in 1947 and 1951 (Nelson *et al.* 2008). As described in Section 1.7, reassortment between multiple lineages of influenza viruses circulating in swine occurred prior to the emergence of the H1N1-2009 pandemic strain (Smith *et al.* 2009). The 'genotype Z' form of HPAI H5N1 which is circulating in Asian poultry also arose as a result of a number of reassortment events with other avian strains (Li *et al.* 2004).

Recombination refers to the exchange of genetic material between segments of RNA. Thus, reassortment is essentially recombination between entire RNA segments and, as such, is sometimes called *intergenic* recombination. In contrast, *intragenic* recombination refers to the exchange of genetic material between parts of gene segments and can be further subdivided into homologous (between RNA segments of the same type) or non-homologous (between different types of RNA segment). Gene segments comprising genetic material from more than one parent virus are known as mosaics. Purported cases of non-homologous intragenic recombination in influenza viruses in the natural world have been documented. For example, Suarez *et al.* (2004) argued that recombination between the HA and nucleoprotein genes of an H7N3 virus, which resulted in a novel HA cleavage site motif, led to a shift in virulence in a flock of poultry in Chile. Suarez *et al.* (2004) found that highly pathogenic viruses from the Chilean H7N3 outbreak possessed a 10 amino-acid insertion at the HA cleavage site which did not conform to the usual HP sequence motifs (see Section 1.6), but shared 100% nucleotide sequence identity with a region of the nucleoprotein gene.

Whilst homologous recombinant influenza viruses can be created in the laboratory, their existence in the natural world is a point of some controversy. In a large-scale study, Obenauer *et al*. (2006) failed to detect homologous recombination in influenza viruses isolated from wild birds and speculated that, if it does exist, it constitutes a rare event. Although some studies claim to have found mosaic influenza viruses in birds (e.g. He *et al*. (2008b)) and swine (e.g. He *et al*. (2008a)), it has been argued that the methods are detecting sequencing artefacts rather than genuine recombinants (Boni *et al.* 2008). For example, Krasnitz *et al*. (2008) observed a substantial overlap between inferred recombinant sequences from the NCBI Influenza Virus Resource and sequences separated from another by a large period of time and an anomalously low rate of evolution. Krasnitz *et al*. (2008) speculated that contamination of virus stock was the most likely cause of these phenomena, rather than a more complicated evolutionary explanation. Homologous recombination is yet to be detected in human influenza viruses (Boni *et al.* 2008).

In contrast to the rapid nature of the evolution driven by reassortment and intragenic recombination, influenza A viruses also evolve via the accumulation of single amino acid substitutions. The nomenclature for amino acid mutations in avian influenza viruses is such that, for example, a mutation from glutamic acid (abbreviated to E) to lysine (K) in the PB2 protein would be written as ‘PB2-E627K’. The purported lack of amino acid substitutions in wild aquatic fowl (Gorman *et al.* 1990a; Gorman *et al.* 1990b), combined with the observation that influenza infection is rarely symptomatic in ducks, led some researchers to conclude that influenza viruses have reached an evolutionary stasis in such hosts and are fully adapted to their natural reservoir (Webster *et al.* 1992). However, a comprehensive study of second codon position substitutions (all of which lead to a different amino acid being coded for) indicated that influenza viruses experience high rates of amino acid substitution across their host range (Chen and Holmes 2006). The evolutionary rate of influenza viruses has been found to accelerate after transmission to a new host species (Ludwig *et al.* 1995; Suzuki and Nei 2002).

1.11 Aims of study

Despite the vast amount of research conducted on avian influenza viruses to date, many questions relating to their evolution remain unanswered. Improved understanding of avian influenza virus evolution could help to minimise the economic impact of influenza outbreaks in poultry and other livestock, as well as the health threat posed to humans. From a human health perspective, it is important to be aware of how avian influenza viruses are evolving in order to identify those which could give rise to a pandemic strain. The need for surveillance studies and analysis of how influenza viruses are circulating in wild and domestic poultry, as well as in other animals, was reinforced by the unheralded emergence of H1N1-2009 (Fraser *et al.* 2009; Smith *et al.* 2009).

The process of reassortment, which creates genetically novel viruses as described in Section 1.10, is well documented in avian influenza, but the rate at which it occurs remains to be quantified (Nelson and Holmes 2007). Furthermore, although reassortment can produce viruses with novel combinations of gene segments, few studies have considered how the genetic interactions between different segments affect their evolution (Rambaut *et al.* 2008). Elucidating genetic interactions, in particular between segments with known functional links, is therefore a key area for influenza research. Formalising links between the evolutionary and epidemiological dynamics of influenza viruses, coupled with the use of intensively sampled outbreak data, may also yield significant benefits in terms of our ability to model and understand avian influenza outbreaks.

This PhD uses phylogenetic and computational techniques to address questions relating to evolutionary change in avian influenza viruses. Firstly, I perform a phylogenetic analysis of all available H7 avian influenza HA sequences in the NCBI database to investigate global evolutionary relationships. In subsequent chapters I perform more detailed molecular and phylogenetic analyses. I investigate the use of stochastic mutational mapping to estimate selective pressures in parts of the H7 HA phylogeny corresponding to lineages associated with different NA subtypes and detect amino acid sites under putative positive selection. I also use Bayesian ancestral state reconstruction methods to investigate patterns and rates of reassortment amongst H7 avian influenza viruses. These methods can also be used to investigate rates of recombination more generally, and I apply them to investigate the history of recombination between different subtypes of HIV-1 group M in the Democratic Republic of the Congo. Finally, I provide a detailed explanation of recent theoretical work on how viral sequence data sampled during an epidemic can be used to estimate epidemiological parameters such as the rate of disease transmission, and discuss the potential for application to avian influenza outbreak and HIV sequence data.

Chapter 2

Methods for analysing viral evolution

2 Methods for analysing viral evolution

This thesis focuses on the evolutionary analysis of viral sequence data, in particular from avian influenza viruses. Many of the methods employed in the thesis are common to more than one chapter and are described here for reference. An overview of frameworks for statistical inference is provided, as is a discussion of the range of methods available for modelling viral evolution and a description of the main software available for this purpose.

2.1. Statistical frameworks

Two main statistical frameworks exist for hypothesis testing, model construction and comparison. This section describes the rationale and some technical details underlying the maximum likelihood and Bayesian paradigms, which are used for phylogenetic and other statistical inferences. Detail is provided on how to interpret results of analyses carried out under these frameworks. A description of the parsimony criterion for developing models and choosing between hypotheses is also presented.

2.1.1 Maximum likelihood

The maximum likelihood (ML) method assesses the compatibility of observed data with the hypothesis that a particular model fits that data. For data D, the likelihood of the data given a model M with parameters θ is denoted $f(D| \theta,M)$, i.e. the probability of the data given the model and associated parameter values. Maximum likelihood estimation refers to the process of finding the values of the parameters under which the probability of observing that data would be maximised for a particular model (i.e. under which the likelihood attains its maximum value). Given independent and identically distributed observed data points $x_1,...,x_N$ and a probability density function *f*(*x*), maximum likelihood determines estimates $\hat{\theta}$ of the parameter *θ* such that:

$$\hat{\theta} = \arg\max\left[\prod_{i=1}^{N} f(x_i \mid \theta, M)\right].$$

2.1.2 Bayesian inference

Whilst frequentist methods such as maximum likelihood consider the parameter *θ* to be a fixed but unknown quantity, *θ* is considered to be a random variable in Bayesian statistics and will thus have a distribution. Bayesian inference also considers a likelihood function analogous to that described in Section 2.1.1, but in addition incorporates existing beliefs about *θ* in the form of a prior distribution. The prior distribution and the likelihood are combined to obtain a posterior distribution for *θ*; the quantities are related by the rule that the posterior is proportional to the product of the prior and the likelihood. A non-informative prior, also known as a diffuse prior, may be used when there is no obvious candidate for a prior (see Hastie *et al.* 2001). The uniform distribution is an example of a non-informative prior.

Bayes' theorem states that, for events A and B:

$$p(B \mid A) = \frac{p(A \mid B)p(B)}{p(A)}.$$

In the above form, Bayes' theorem can be used for calculating conditional probabilities. For parameters *θ*, conditioned on a model *M*, with observed data *D*, one can write an expression involving density functions:

$$f(\theta \mid D, M) = \frac{f(D \mid \theta, M) f(\theta \mid M)}{f(D \mid M)}.$$

$f(\theta|D,M)$ is the posterior probability distribution for the parameters given the data and model, $f(D|\theta,M)$ is the likelihood of the data given the parameters and $f(\theta|M)$ is the prior distribution representing existing beliefs about the parameters.

When Bayesian inference has been used to approximate a posterior distribution (e.g. using MCMC – see Section 2.4), uncertainty in parameter estimates can be reported using Bayesian credible intervals. A 100*(1-α)% credible interval for a posterior distribution for θ is any interval [a, b] in the domain of the distribution such that the posterior probability of θ lying between a and b is 1 – α. The highest posterior density (HPD) interval is the narrowest such credible interval. The upper and lower limits of HPD intervals are used in Bayesian hypothesis testing. For example, non-overlapping HPD intervals may be interpreted as associated with parameters which are significantly different in value. In addition, if zero lies outside of an HPD interval then this may be interpreted as significant evidence that a parameter is non-zero. In practice, a 100*(1-α)% HPD interval can be calculated from observations sampled from a posterior distribution as the narrowest interval containing 100*(1-α)% of the data. It must be noted that HPD intervals are defined differently to the confidence intervals used in frequentist statistics such as maximum likelihood. A 100*(1-α) % confidence interval for a parameter is constructed such that, if the confidence interval was calculated in the same manner from multiple independent sets of observations from a distribution then the interval would contain the true value of the parameter 100*(1-α) % of the time.

2.1.3 The parsimony criterion

Although not a formal statistical technique, the parsimony criterion may be used to choose between models or evolutionary hypotheses. Parsimony is based upon a preference for the simplest explanation which is compatible with the data. For example, the preferred phylogenetic tree under parsimony is one which requires the minimum number of evolutionary changes to reconcile the sequences.

2.2 Model selection

This section describes methods for choosing between different models which have been fitted to observed data.

2.2.1 Likelihood ratio tests

Models may differ in the number of parameters of which they comprise. Where a simpler model may be obtained by collapsing a more complex model, the models are described as nested. If one model is not simply a special case of the other then the models are said to be non-nested. A likelihood ratio test (LRT) may be used to compare the fit of models which have different numbers of parameters. The test statistic for a likelihood ratio test of model M_2 model against a simpler (null) model M_1 is given by:

$$-2[ln(\mathrm{L}(D|M_1)) - ln(\mathrm{L}(D|M_2))].$$

Note that 'L' is the standard statistical notation for a likelihood in a non-Bayesian setting and is used here be consistent with the literature on test-statistics; the expression $\mathrm{L}(D|M)$ is equivalent to $f(D|M)$ presented in previous and subsequent sections.

For nested models, the likelihood ratio test statistic is distributed approximately chi-squared, with number of degrees of freedom equal to the difference in the number of parameters between the two models. If the models are non-nested then parametric bootstrapping can be used to assess the significance of the LRT (see Goldman (1993a); Goldman (1993b) and Huelsenbeck *et al*. (1996) for discussion). Parametric bootstrapping involves generating a null distribution for the LRT statistic by simulating a large number of datasets under the null model and observing where the LRT score for the original data falls in the null distribution for the LRT statistic.

Generally, if less than 5% of the simulated LRT scores are at least as extreme as the alternative LRT score then this is taken to be significant evidence for rejecting the null hypothesis.

2.2.2 Bayes factor tests

In a Bayesian analysis, models can be selected by comparing their marginal likelihoods. The marginal likelihood for a model *M* is the probability of the observed data *D*, given the model *M*, averaged over the model parameters *θ*:

$$f(D\,|\,M) = \int f(D\,|\,\theta, M)\, f(\theta\,|\,M)\, d\theta\,.$$

$f(D|M)$ is the denominator in the Bayes theorem expression for parameters *θ* conditioned on a model *M*, with data *D* (see Section 2.1.2). For models M_1 and M_2, the ratio of the marginal likelihoods of the two models, $f(D|M_1)/f(D|M_2)$, known as the Bayes factor, can be used to make comparisons between the models.

A Bayes factor which is deemed to be significantly greater than 1, i.e. for which $f(D|M_1)$ is significantly greater than $f(D|M_2)$, is interpreted as statistical support for choosing model M_1 over M_2. Conversely, $f(D|M_2)$ significantly greater than $f(D|M_1)$ provides support for M_2 over M_1. Although formal rules do not exist, guidelines have been suggested for the interpretation of Bayes factors, for example by Jeffreys (1961) and Kass and Raftery (1995). Typically, a Bayes factor of greater than 20, or the natural logarithm of a Bayes factor being greater than 3, is taken as significant support for one model over another. There is no requirement for models being compared using Bayes factor tests to be nested.

In order to calculate Bayes factors, the marginal likelihoods of the models under comparison are evaluated using Bayesian MCMC (see Section 2.4) over both models, known as reversible jump MCMC. The Tracer software (Drummond and Rambaut 2007) can be used to analyse Bayesian MCMC output from phylogenetics software such as BEAST and MrBayes (see Section 2.6.3 for a discussion of these software). As reversible jump MCMC was not implemented in BEAST at the time that the studies in this thesis were performed, the importance sampling method described by Newton *et al*. (1994) was used to approximate marginal likelihoods and calculate approximate Bayes factors by processing the BEAST output in Tracer.

2.2.3 Akaike information criterion (AIC)

Whilst LRTs and Bayes factor tests can only be used to test between pairs of hypotheses, information criteria can be used to compare multiple models $\mathrm{M}_1,\ldots,M_N$ simultaneously. The models under comparison can be nested or non-nested. The Akaike information criterion (AIC) (Akaike 1974) is a penalised log-likelihood which addresses the trade-off between the fit of a model to the data and the complexity of the model by requiring that the addition of parameters improves the log-likelihood of the model by a certain amount. The AIC score for a model M_i is given by:

$$\mathrm{AIC}(M_i) = -2\ln(L_i) + 2\,N_i\,,$$

N_i is the number of free parameters in model M_i. The maximum likelihood score, L_i, for the data under model M_i, can be computed and compared for models M_1 to M_N. The preferred model would be the one with the lowest AIC score.

2.2.4 Bayesian information criterion (BIC)

The Bayesian information criterion (BIC) can also be implemented when maximum likelihood is used to fit models to data. BIC was developed by Schwartz (1979). The BIC score is given by:

$$\mathrm{BIC}(M_i) = -2\ln(L_i) + N_i \ln(n)\,.$$

BIC is thus similar to AIC, but in the penalty term the number of parameters is multiplied by ln(n) rather than 2. This generally leads to BIC being more stringent in terms of the number of parameters, and choosing simpler models than AIC. The BIC score is related to the marginal likelihood $f(D|M)$, which is calculated for each model in a Bayes factor comparison (see Ripley (1996)). Choosing a model with the smallest BIC score is approximately equivalent to choosing the model with the highest posterior probability (for more details see Hastie *et al*. (2001)).

2.3 Markov processes

Markov processes are used in evolutionary biology to model transitions between discrete states. For example, Markov processes are used to model substitution of nucleotides, amino acids and codons in genetic sequence data, as well as for Bayesian inference of ancestral states along phylogenies. The set of possible states in a Markov process is known as the state space. A Markov process is characterised by transition rates between pairs of states, usually denoted q_{ij} to refer to the rate of transition from state i to state j. The matrix of transition rates is referred to as the **Q**-matrix. Markov processes are the continuous-time generalisations of Markov chains and are sometimes known as continuous-time Markov chains (CTMC).

A Markov process satisfies the Markov property of being ‘memoryless’. This means that the future state depends solely on the current state and is independent of the history of previous states. For past times $t_1, t_2, \ldots, t_n$, present time t and future time s, this can be formalised as $p(x_s|x_t, x_{t_n}, ..., \mathrm{x}_{t_1}) = p(x_s|x_t)$, where $s > t > t_\mathrm{n} > \ldots > t_1$. In a Markov process, the waiting times between jumps are exponentially distributed. From state i, the waiting time until the first jump is exponentially distributed with rate parameter $-q_{ii}$. A Markov process is said to be time-homogeneous if $p(x_s|x_t) = p(x_{s+n}|x_{t+n})$, i.e. if the transition probabilities are equal for time intervals $(s - t)$ of the same length, anywhere in the process.

2.4 Markov chain Monte Carlo sampling

The method of Markov chain Monte Carlo (MCMC) sampling is implemented in many packages for phylogenetic and other statistical inferences, and is used throughout this thesis. MCMC is a technique for approximate-sampling from a target distribution, using a Markov chain constructed to have a stationary distribution (the distribution when the chain is at equilibrium) with the same properties as the target distribution from which one endeavours to sample. After convergence to the stationary distribution, the proportion of time that the Markov chain spends in each state is proportional to the probability of that state in the target distribution. MCMC is widely used in Bayesian inference, where the target distribution would be the posterior distribution for the parameter being estimated. Algorithms such as Metropolis-Hastings (Metropolis *et al.* 1953; Hastings 1970) are used to obtain random samples from a distribution using MCMC, by proposing moves to different states and accepting them according to some probability.

A burn-in period is usually required to obtain convergence to the stationary distribution using MCMC. A diagnostic plot of the chain trace, which shows the likelihood of each model against the sample generation, is widely used to assess convergence. After the burn-in period, sampling from the Markov chain is undertaken at periodic intervals as specified by the user. The sampling interval should be sufficiently wide as to ensure that the samples are not autocorrelated. The effective sample size (ESS) of a post-burn-in sample, which is given by the post-burn-in chain length divided by the average number of states in the chain by which two samples must be separated for them to be uncorrelated, provides an estimate of the number effectively independent samples from the posterior distribution to which the MCMC is equivalent. Comparing MCMC output from multiple independent runs can indicate whether chains are converging on the same distribution.

Although MCMC aims to sample across a target distribution, in phylogenetic inference the distribution of trees may have multiple local peaks associated with different topologies, which are difficult to move between. Metropolis coupling (Geyer 1991) is used to improve the mixing of the chain in Bayesian phylogenetics software such as BEAST and MrBayes. Metropolis coupling involves running several chains in parallel. One chain, referred to as the cold chain, samples from the target distribution of interest. The cold chain is coupled with multiple heated chains h_i, whose stationary distributions are obtained by raising the stationary distribution of the cold chain to a power $1/r_i$ (with $r_i > 1$ for all i). This coupling results in 'flatter' distributions in which is it easier to move between the local peaks. At regular intervals, swaps between the states of two randomly chosen chains are proposed and accepted in a probabilistic manner similar to that used in Metropolis-Hastings sampling. Involvement of the cold chain in a swapping event can lead to jumps between local peaks and a more thorough exploration of the target distribution. Output from the heated chain is discarded at the end of the run, and only output from the cold chain is used.

2.5 Modelling sequence evolution

Modelling genetic sequence evolution is a prerequisite for computing genetic distances between sequences. Amongst other things, genetic distances can be used in constructing phylogenetic trees from sequence alignments. In this section, the framework for constructing nucleotide substitution models as a Markov process is outlined, along with examples of some of the widely used models for nucleotide, amino acid and protein evolution.

2.5.1 Nucleotide substitution models

Nucleotide substitution models are statistical models of the process of substitution of the bases A, C, T and G at sites in a nucleotide sequence. Nucleotide substitution models are constructed as Markov processes; for a derivation, see Strimmer and von Haeseler (2009). The Markov property is satisfied for nucleotide substitution models since, for each site, the rate of change from one base to another is assumed to depend solely on the identity of the current nucleotide at that site and to be independent of all previous nucleotide identities at the site.

A Markov chain with transition rate matrix $\mathbf{Q}=\{q_{\mathrm{ij}}\}$ and stationary distribution $\mathit{\Pi}$ is said to be time-reversible if and only if $\pi_{\mathrm{i}}q_{\mathrm{ij}}=\pi_{\mathrm{j}}q_{\mathrm{ji}}$ for all i and j (with $i \neq j$). For mathematical convenience, most nucleotide substitution models are time-reversible. In the general time-reversible (GTR) model (Tavaré 1986), substitution rates are assumed to be constant over time and the relative frequencies of the bases A, C, G and T (denoted π_{A}, π_{C}, π_{G}, and π_{T} respectively, with $\pi_{\mathrm{A}} + \pi_{\mathrm{C}} + \pi_{\mathrm{G}} + \pi_{\mathrm{T}} = 1$) are at equilibrium. Instantaneous rates of change between pairs of bases are given by the product of f_{ij} (the rate of substitution of the old base (i) for the new base (j) relative to all other substitutions) and the equilibrium frequency of the new base (π_j). For example, the instantaneous rate of change from A to C would be given by given by $f_{\mathrm{AC}}* \pi_{\mathrm{C}}$, where f_{AC} is the relative rate parameter for substitution from A to C. A rate matrix, known as the **Q** matrix, is constructed whose off-diagonal entries q_{ij} are given by $f_{ij}* \pi_j$ and whose diagonal entries q_{ii} are chosen so that the sum of each row is zero. Under the conditions of time-reversibility, $f_{\mathrm{ij}}=f_{\mathrm{ji}}$ in GTR model for all i and j (with $i \neq j$).

The entries of the matrix $\mathbf{P}(t) = \exp(\mathbf{Q}t)$ are finite-time transition probabilities, with $p_{ij}(t)$ denoting the probability that a substitution from base i to base j takes place at a given site along a branch of length t. The transition probabilities can be used to calculate genetic distances between pairs of sequences in an alignment and thus to construct a phylogenetic tree (see Section 2.6). The notion of the expected distance between pairs of sequences enables the likelihood of a given topology to be calculated (Felsenstein 1981).

The GTR model allows different equilibrium base frequencies for each nucleotide and different rates of transition between pairs of bases under the condition of time-reversibility. All time-reversible nucleotide substitution models are essentially special cases of the GTR model. The simplest nucleotide substitution model was introduced by Jukes and Cantor (1969) and assumes equal base frequencies at equilibrium and that each nucleotide is equally likely to replace another. Later models extended the Jukes-Cantor model to allow for different rates of transition and transversion, for example the K2P model of Kimura (1980). The HKY model of Hasegawa *et al*. (1985) also allowed for different rates of transition and transversion, at the same time as relaxing the assumption of equal base frequencies. Rate heterogeneity across sites can be modelled by a gamma distribution with varying shape parameter, alpha (Yang 1994; Yang *et al.* 1994), and a proportion of invariant sites may also be included.

Whilst more complex substitution models may provide a more accurate description of biological reality, they are also more computationally intensive. For example, the **Q-**matrix for the HKY model may be exponentiated analytically, whereas the **Q-**matrix for the GTR model cannot. Furthermore, more complex models can lead to larger variances when branch lengths are being estimated from genetic distances between pairs of sequences, in particular when the sequences being compared are short (less than 1000 nucleotides) and the assumption that the number of sites being compared is infinite is clearly violated (Strimmer and Von Haeseler 2003). The fit of different substitution models relative to the number of parameters in the model may be assessed using methods for model selection. Software available for this purpose includes ModelTest (Posada and Crandall 1998) for nucleotide substitution model selection using LRT values computed using PAUP*, as well as model selection options for nucleotide, codon and amino acid data implemented in the HyPhy software (Kosakovsky Pond *et al.* 2005).

2.5.2 Amino acid and codon substitution models

Empirical models of amino-acid substitution (e.g. Dang *et al*. (2010) for influenza; Nickle *et al*. (2007) for HIV) and theoretical models of codon substitution (Goldman and Yang 1994; Muse and Gaut 1994) are also constructed as Markov processes. By explicitly incorporating information about the genetic code, codon models may better reflect biological reality than models of nucleotide substitution; however, running full codon models is computationally intensive. Shapiro *et al*. (2006) argued that partitioning substitution rates into third codon position rates and combined first and second position substitution rates improved the fit of nucleotide models to protein coding data and was more computationally efficient than using full codon models. Partitioning substitution rates according to the SRD06 model of Shapiro *et al*. (2006) reflects the increased tendency for nucleotide changes at the first and second positions to change the amino acid coded for, compared to third position substitutions which are more likely to leave the amino acid unchanged. The SRD06 model employs the HKY model of nucleotide substitution, allowing base frequencies to be estimated from the data and gamma distributed rate heterogeneity across sites. Under the SRD06 model, substitution rates, the transition-transversion ratio and rate heterogeneity across sites may be unlinked between the third codon position and the first and second codon positions.

2.6 Phylogenetic inference

Phylogenetic trees are used to show evolutionary relationships between groups of molecular sequences. More closely related sequences cluster together more closely in a tree than those which are less closely related. This section provides a brief description of the main frameworks for inferring phylogenies. Bayesian methods for phylogenetic reconstruction are considered, in particular those which incorporate sample date information to produce a tree with an explicit time-scale. Such methods have found widespread application in the analysis of viral sequence data and are implemented extensively in this thesis.

2.6.1 Constructing phylogenies

Phylogenetic trees can be constructed under parsimony, maximum likelihood and Bayesian frameworks, as well as using distance-based methods. Parsimony, ML and Bayesian methods infer phylogenies directly from the discrete character information in the sequences, whereas distance methods use scores between all pairs of sequences based upon the number of substitutions inferred to have taken place under an evolutionary model. ML and Bayesian methods use an explicit substitution model, whereas parsimony seeks to construct a topology based upon the minimum amount of evolutionary change necessary to reconcile the sequence data. In this thesis, the MEGA software (Tamura *et al.* 2011) was used to conduct distance-based (neighbour-joining) phylogenetic analyses, and PHYML (Guindon *et al.* 2010) was used to construct ML trees. MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) and BEAST (Drummond and Rambaut 2007) were used for Bayesian inference of phylogenies.

Phylogenetic inference methods can broadly be classified as criterion-based or algorithmic. Algorithmic methods, such as neighbour-joining (NJ) (Saitou and Nei 1987), cluster taxa according to a pre-defined set of rules, whereas criterion-based methods such as parsimony, ML, and certain distance-methods such as least squares, search for the best tree according to some criterion. Algorithmic methods will return a single tree based upon a series of operations, whereas the search for an optimal tree in criterion-based methods means that multiple trees are considered and the ‘best’ tree is reported.

Criterion-based methods are more computationally intensive than algorithms such as NJ. The number of possible tree structures for *n* sequences increases super-exponentially with *n*, thus it would be impossible to compare all possible tree structures for more than 10 taxa (Cavalli-Sforza and Edwards 1967). Heuristic methods are therefore employed to search the space of possible trees, but are not guaranteed to find the best tree and may become stuck at a local optimum. When a fast method for inferring relationships between taxa is required, NJ trees have been shown to provide a good approximation to the minimum evolution tree (the tree with the smallest least-squares measure of observed distances between sequences and the distances predicted by the tree), particularly for large datasets (Strimmer and von Haeseler 1996).

2.6.2 Maximum likelihood phylogenetics

ML phylogenies are constructed by determining the topology, branch lengths and substitution model parameters which maximise the probability of observing the sequence data upon which the tree is being inferred. This involves calculating ML estimates of branch lengths and parameters for different topologies, then selecting the phylogeny with the highest overall likelihood. Felsenstein's pruning algorithm (Felsenstein 1973; Felsenstein 1981) can be used to calculate the likelihood of a particular phylogeny for a given sequence alignment. Branch lengths which maximise the likelihood can be calculated for each topology using numerical methods. Parameters of the substitution model, such as the alpha parameter for rate heterogeneity across sites, may be estimated on an initial NJ tree prior to the ML tree search, rather than concomitantly with the tree which maximises the likelihood function. Heuristic methods for tree-rearrangement which may be used to search over the space of phylogenies include nearest-neighbor interchange, sub-tree pruning and regrafting, and tree bisection and reconnection (see Felsenstein (2004) for more details).

2.6.3 Bayesian phylogenetic inference

Phylogenetic inference under the Bayesian framework involves computing the posterior distribution of evolutionary parameters, given aligned molecular sequence data. The parameters being inferred include a topology and branch lengths, but could also include the parameters of a substitution model and, for genealogy-based inference using coalescent processes, population genetic parameters. The MrBayes software focuses on phylogenetic inference under a Bayesian framework, whilst BEAST (Bayesian Evolutionary Analysis by Sampling Trees) provides a coalescent-based framework for concomitant inference of phylogenetic and population genetic parameters.

Both MrBayes and BEAST use MCMC to sample from the posterior distribution of evolutionary parameters. The collection of post-burn-in samples obtained after convergence to the stationary distribution is an approximation to the posterior distribution of the parameter under consideration. Sampling in this manner accommodates uncertainty in topologies, branch lengths and the parameters of the substitution model, whereas NJ and ML respectively provide a single tree according to a set of rules or the 'best' tree under some criterion. Indeed, maximum likelihood estimates of evolutionary parameters based upon a single tree have been criticised for having artificially low confidence intervals as they do not account for phylogenetic uncertainty (Nielsen 2002). However, ML methods do have the advantage that they can be used to directly compare alternative hypotheses, for example in forensics (Holmes *et al.* 1993). Posterior samples of phylogenies may be summarised in a variety of manners (see Section 2.9) and HPD intervals for parameters can be obtained to enable hypothesis testing.

2.7 Relaxed clock phylogenetics

The earliest phylogenetic methods assumed a constant rate of molecular evolution across all lineages, however diverse, and within each lineage over time. The model of rate-constancy across the tree is known as the strict molecular clock (Zuckerkandl and Pauling 1965). However, it is widely believed that the assumption of a strict clock can be violated in nature, for example due to changes in selective pressure (Kimura 1986) which may be particularly relevant to emerging viral diseases moving into different hosts, or subject to pressures such as vaccination or the use of anti-viral drugs. The incorrect application of a strict clock can lead to error in assigning phylogenetic relationships and estimating divergence times (e.g. Ayala (1997); Ho and Jermiin (2004)).

Felsenstein's (1981) alternative to the strict molecular clock invokes the assumption that evolutionary rates are independent for each branch of the tree. However, this model is unable to separate the effect of the substitution rate from evolutionary time in causing observed differences between sequences, and an outgroup sequence or a non-reversible model of nucleotide substitution is required to root the tree and infer the direction of evolution. It is not possible to estimate substitution rates in terms of an explicit timescale under the unrooted model. Many of the major phylogenetic software packages, including MrBayes, PhyML and MEGA, provide a choice of only the strict molecular clock or Felsenstein's unrooted model and are thus limited.

Drummond *et al*. (2006) implemented a new class of phylogenetic models in the BEAST software, which used Bayesian MCMC to provide joint estimates of divergence dates and phylogenies. These models are known as relaxed clock models, and provide an intermediate between the strict clock and Felsenstein's unrooted model. In relaxed clock phylogenetics, evolutionary rates along branches are sampled from a user-specified distribution (usually lognormal). Relaxed phylogenetic models may be uncorrelated or autocorrelated. In autocorrelated models, rates vary across the tree such that the rate along a given branch depends on the rate along the parental branch (the branch immediately preceding it towards the root of the tree). For uncorrelated models the branch rate does not *a priori* depend on the rate of the parental branch.

In BEAST, a Bayes factor test between relaxed and strict clock models can be used as evidence to reject the strict molecular clock. Drummond *et al*. (2006) found that, whilst erroneously assuming a strict molecular clock in the presence of rate heterogeneity across a tree could confound phylogenetic inference, uncorrelated relaxed clock models still performed reasonably on clock-like data. Drummond *et al*. (2006) found no evidence of autocorrelation between parent and child branches for viral datasets including influenza A, and the uncorrelated relaxed lognormal (ucln) clock model is currently implemented in BEAST as an alternative to the strict clock.

2.8 Incorporating sample date information

Under both strict and relaxed molecular clock models, separation of substitution rate from evolutionary time is only possible in the presence of external calibration data (discussed by Drummond *et al*. 2006). In BEAST, such information often comes from the sampling dates of sequences sampled over a sufficiently large time-span. In the absence of adequate time-stamped data, a strong prior distribution for the substitution rate is required. When no prior information for the substitution rate is available, the substation rate may be fixed to 1, so that the branch lengths of the tree are in units of substitutions per site, rather than units of years.

Viruses such as HIV or influenza, which have a mutation rate so high that their evolutionary and ecological dynamics may be observed on the same timescale, are known as measurably evolving (Drummond *et al.* 2003). Phylogenetic analysis of sequence data from measurably evolving populations which have been sampled at sufficiently diverse time-points can provide an insight into the evolutionary dynamics and demographic history of a population (Drummond *et al.* 2002).

2.9 Summarising sets of phylogeny samples

The method of bootstrapping has been applied in phylogenetics (Felsenstein 1985a) to allow uncertainty in phylogenetic reconstructions to be assessed. Bootstrapping is a technique for approximating an unknown or analytically intractable statistical distribution by resampling from the original dataset (Efron and Gong 1983). For an alignment of sequences where the rows represent different taxa and the columns are sites along the genome, columns of sites are randomly sampled with replacement to create a new alignment of the same size as the original. New alignments sampled in this way are known as bootstrap samples. The resampling process is carried out many times (typically 1000) and phylogenies (known as bootstrap trees) are constructed from the bootstrap samples. The proportion of the bootstrap phylogenies in which particular clades appear is then reported. Theoretically, the bootstrap phylogeny samples should approximate the variance of a set of trees obtained using a set of alignments constructed using a set of new sites each time. Bootstrapping is often performed on individual trees obtained using maximum likelihood or distance-based methods.

It is often desirable to summarise tree samples from a Bayesian phylogenetic analysis by reporting posterior probabilities of clades. Post-MCMC analysis in the MrBayes software enables a 50% majority rule consensus tree (Margush and McMorris 1981) or a fully resolved majority rule consensus tree to be constructed. A 50% majority rule consensus tree contains all clades appearing in at least half of the tree samples, whilst clades in a fully resolved majority rule consensus tree are chosen in order of decreasing posterior probability, excluding clades incompatible with previously selected clusters.

Majority rule methods for summarising phylogenies have the disadvantage that the consensus tree may not actually have been sampled in the MCMC and thus may not best reflect biological reality. An alternative is to choose a maximum *a posteriori* tree from the set of samples. This could be the tree in which the sum of the posterior probabilities of clades is maximised (the maximum clade credibility, or MCC, tree), or the tree in which the product of the posterior probabilities of clades is maximised (the maximum credibility tree). Alternatively, the smallest set of trees which account for at least *x*% of the total posterior probability in the sample under these criteria may be reported. Such sets are known as *x*% credible sets.

In this thesis, the TreeAnnotator software (http://beast.bio.ed.ac.uk/TreeAnnotator) was used to construct summary phylogenies from tree samples obtained through analysis with BEAST. FigTree (http://tree.bio.ed.ac.uk/software/figtree) was used to visualise and manipulate the phylogenies presented.

2.10 Mapping discrete traits onto phylogenies

It is often possible to reconstruct trait evolution which has occurred alongside sequence evolution in a measurably evolving population (Holmes 2004). For example, phylogeographic methods for the concomitant study of spatial diffusion and sequence divergence have been used to investigate the ancestral location and spread of highly pathogenic H5N1 avian influenza (Lemey *et al.* 2009) and other emerging viral diseases such as dengue (Raghwani *et al.* 2011). Characters corresponding to multiple traits can be mapped concomitantly onto phylogenies, and correlations in the histories of different traits may be detected (e.g. Maddison (1990); Pagel and Meade (2006)). Reconstruction of ancestral states along a phylogeny can be carried out under parsimony, maximum likelihood and Bayesian frameworks.

2.10.1 Parsimony reconstruction of ancestral states

Early attempts at ancestral state inference invoked the parsimony criterion, for example the two-pass parsimony algorithm (see Fitch (1971); Swofford and Maddison (1987)) which is implemented in the MacClade software (Maddison and Maddison 1992). The two-pass parsimony algorithm involves firstly working from the tips of the tree towards the root (the downward pass) to determine the possible ancestral states for each node, then proceeding back from the root towards the tips of the tree (the upward pass) to optimise the assignment of ancestral states under the parsimony criterion. The following rules are applied for the downward pass: if the intersection of the set of states at descendant nodes is non-empty, then the set of shared states is assigned to the ancestor; otherwise, the intersection of the sets of states at the descendent nodes is the empty set and the union of the states at these nodes is assigned to the ancestral node. In the upward pass, if the state of the parent node is in the set of possible child nodes then the child node is chosen to have the same state as the parent. In cases where more than one parsimonious explanation of the ancestral process exists, the upward pass stage can be performed multiple times after the downward pass has been carried out, in order to enumerate all compatible parsimonious reconstructions.

The parsimony approach to ancestral state inference has been criticised for a number of reasons. For example, minimising the number of changes may not attain the most biologically relevant model for rapidly evolving organisms such as viruses (Cunningham *et al.* 1998). In addition, the parsimony method does not account for uncertainty in the phylogenetic reconstruction or the process for which the ancestral states are being inferred (Nielsen 2001).

2.10.2 Maximum likelihood ancestral state reconstruction

Maximum likelihood approaches to reconstructing ancestral states for discrete characters, such as those of Pagel (1994) and Schluter *et al*. (1997), have modelled transition probabilities between states as a Markov process. ML methods remove the need to exclude non-parsimonious mappings and, by using an explicit probabilistic framework, provide a more rigorous method than parsimony for testing evolutionary hypotheses about character evolution (Maddison 1995). Unlike parsimony, ML methods account for branch lengths in calculating the probabilities of ancestral states, reflecting the higher probability of observing a change over a longer period of time compared to over a short period of time.

2.10.3 Bayesian methods for ancestral state reconstruction

Bayesian methods for ancestral state inference explicitly account for uncertainty in the phylogeny and the substitution model parameters by considering multiple MCMC samples (Nielsen 2001; Nielsen 2002). Like the ML methods outlined above, Bayesian approaches also model trait evolution using continuous-time Markov chains and do not exclude non-parsimonious mappings.

The Bayesian stochastic mutational mapping method of Nielsen (2001) explicitly accounts for uncertainty in the mutational history, by sampling multiple mutational mappings for each phylogeny sample. Originally, stochastic mutational mapping was developed for mapping nucleotide mutations onto phylogenies. The method was extended by Huelsenbeck *et al.* (2003) for mapping morphological trait characters along phylogenies, and testing for correlations in the evolutionary history of such traits. The process of constructing a stochastic mutational map under the method of Nielsen is outlined below. The description is for nucleotide states, but an analogous process is used in the methods of Huelsenbeck *et al.* (2003) for mapping other discrete characters onto a phylogeny, where transitions between states can be modelled as a Markov process. Stochastic mutational mapping and stochastic character mapping are implemented in the SIMMAP software (Bollback 2006).

In Bayesian stochastic mutational mapping, mutational histories are sampled from the posterior distribution of mappings, given the observed nucleotide data. Mutational histories may be inferred for a given nucleotide site as follows. Firstly, the fractional likelihoods for nucleotides A, C, T and G are calculated at each node in the rooted phylogeny using the method of Felsenstein (1981). Next, the ancestral state at the root of the tree is simulated using the fractional likelihoods for nucleotides at the root. Ancestral states of nodes further towards the tip of the tree may then be sampled recursively, conditioned on the data and the states at all previous nodes. Finally, mutational histories are simulated for all lineages (between pairs of parent and child nodes) by modelling the substitution process from an ancestral node using a continuous-time Markov chain. For a dataset *D*, an infinite number of possible mutational histories exist and each mapping *A* has an associated probability which can be evaluated as:

$$P(A \mid D) = \frac{P(A, D)}{P(D)}.$$

Mappings are then sampled in proportion to their posterior probability and nucleotide transitions may be visualised along the tree (see Chapter 4, Figure 4.1). Stochastic mutational mapping is implemented in Chapter 4, where an extension of the method of Nielsen (2001) is used to estimate selective pressure along lineagess.

Another Bayesian ancestral trait mapping method was provided by Lemey *et al*. (2009) and is implemented in BEAST. The method of Lemey *et al*. (2009) was originally developed as a spatial diffusion model for phylogeographic inference. As with other probabilistic methods for ancestral state inference, such as stochastic mutational mapping (Nielsen 2001), Lemey *et al*. model transitions between states ancestral states as a Markov process. The discrete trait mapping procedure of Lemey *et al*. (2009) reports posterior probabilities of the inferred discrete trait state at each node of the tree, as well as relative rates of transition between pairs of states (see Section 5.4.2 for more detail). Lemey *et al*. (2009) also employ the technique of Bayesian stochastic search variable selection (BSSVS) to identify a parsimonious description of the diffusion process (i.e. the model with the least number non-zero of transition rates) for the data. Under BSSVS, individual transition rates between pairs of states are switched off (set to zero) or switched on (non-zero) at different steps in the Markov chain. The proportion of the time a rate is switched on or off in the MCMC chain is then considered in Bayes factor testing for significantly non-zero transition rates. The method of Lemey *et al*. (2009) is used in Chapter 5, and is described in more detail in Section 5.4.2.

Given a CTMC for discrete trait transition, Minin and Suchard (2008a) obtained a closed-form analytical solution for the moments of the discrete trait counting process, whereby transitions (‘Markov jumps’) are enumerated along a branch of a phylogeny. The extension of Minin and Suchard (2008b) allowed the timings of the Markov jumps to be tracked, so that the length of time spent in each state along a branch (the ‘Markov rewards’) could be recorded. Minin and Suchard (2008b) also extended their single-branch calculations to discrete trait transitions across the whole tree. Since the method of Minin and Suchard is simulation-free (whereas the stochastic mutational mapping method of Nielsen (2001) requires character histories to be simulated, and discarded unless they are compatible with the data) it is more computationally efficient than simulation-based methods. The method of Minin and

Suchard has been used for inferring synonymous and non-synonymous changes along phylogenies (O'Brien *et al*. 2009). Markov jumps counting has been implemented in BEAST (Talbi *et al*. 2009), where it is used in conjunction with a CTMC inferred under the methods of Lemey *et al*. (2009) (without the use of BSSVS). In this thesis, Markov jumps counting was used in Chapter 5 and Chapter 6.

The discrete trait mapping method of Lemey *et al*. (2009) has been extended to allow ancestral state mapping using continuous trait data, such as latitudes and longitudes, as has the Bayesian trait mapping software BayesTraits (Pagel 1999). The BaTS software (Parker *et al*. 2008) allows the user to assess the extent to which closely related taxa share a particular discrete trait state, for a set of Bayesian phylogeny samples. BaTS can be used to perform a test for association between the phylogeny and the trait, by comparing the observed distribution of states at the tips with a null distribution under which states are randomly distributed across the tips of the tree.

2.11 Estimating selective pressure

2.11.1 d_N, d_S and detecting selection

Because the genetic code is degenerate, many single nucleotide substitutions leave the amino acid coded for unaltered and are known as synonymous, or 'silent'. Single nucleotide changes which result in a different amino acid being coded for are referred to as non-synonymous. By comparing the number of non-synonymous substitutions per non-synonymous site (d_{N}) to the number of synonymous substitutions per synonymous site (d_{S}) (Miyata and Yasunaga 1980), it is possible to make inferences about the nature of the selective forces acting upon a DNA sequence. The scaling 'per synonymous site' or 'per non-synonymous site' in d_{N} and d_{S} accounts for the structure of the genetic code, specifically variation in the number of synonymous changes and non-synonymous changes from different codons.

Synonymous substitutions are often assumed to be selectively neutral, however this may not always be the case. For example, codon usage bias (greater transcription efficiency for particular codons compared to others which code for the same amino acid) can lead to departure from neutrality for synonymous changes (for review, see Hershberg and Petrov (2008) or Sharp, Emery and Zeng (2010)). In addition, the presence of an RNA secondary-structure may lead to suppression of synonymous codon changes (Simmonds and Smith 1999). A potential RNA secondary structure across all eight influenza segments has recently been reported (Moss *et al.* 2011). Overlapping reading frames can also result in synonymous changes being non-neutral, if they lead to non-synonymous changes in an alternative reading frame. Under selective neutrality, rates of non-synonymous and synonymous change would be expected to be equal, hence there would be an expectation that d_N/d_S=1. However, under selective pressure, non-synonymous changes may accumulate at a different rate to non-synonymous substitutions due to a change in fitness associated with a different amino acid. Departures from selective neutrality may therefore be detected by a d_N /d_S ratio which is significantly different from 1. Positive selection refers to the increase in frequency of beneficial mutations in a population towards fixation and is indicated by d_N /d_S >1. Negative selection refers to the reduction in frequency of deleterious mutations towards extinction and has a signature of d_N /d_S < 1. In viruses such as influenza, regions of the gene associated with antigenicity and the evasion of host immune response have been found to be under positive selection (e.g. Suzuki and Gojobori (1999)). The d_N/d_S ratio (also referred to in the literature as $\omega=\alpha/\beta$, or Ka/Ks) can be estimated across a gene, at specific amino sites in an alignment of protein coding sequences and along branches of a phylogeny. Different approaches for calculating d_N /d_S are outlined below.

2.11.2 Methods for calculating d_N/d_S along a gene

The earliest methods for estimating d_N /d_S involved calculating substitution rates for a gene or a region of a gene, so that the d_N /d_S estimate was averaged over several amino acid sites. These approaches can be divided into distance-based and maximum likelihood methods. Distance methods involve obtaining estimates of d_{S} and d_{N} for pairs of nucleotide sequences by dividing the respective numbers of synonymous and non-synonymous differences between the sequences by the numbers of synonymous and non-synonymous sites per sequence. Many distance-based methods exist for estimating $d_{\mathrm{N}} / d_{\mathrm{S}}$, for example those of Miyata and Yasunaga (1980), Perler *et al*. (1980), Li *et al*. (1985), Nei and Gojobori (1986) and Zhang *et al*. (1998).

Methods for estimating d_N and d_S have taken alternative approaches to estimating the number of synonymous and non-synonymous sites. Early methods such as that of Li *et al*. (1985) used a matrix derived from large datasets to establish the likely paths between codons. However, the method of Nei and Gojobori (1986) is now widely used for estimating d_{N} and d_{S}, and bases the estimate on the data analysed as follows. The Nei and Gojobori method involves first computing the number of synonymous (*s*) and non-synonymous (*n*) sites for each codon. If f_i is the fraction of synonymous changes at the i^{th} position of a given codon, then:

$$s = \sum_{i=1}^{3} f_i \text{ and } n = 3 - s.$$

For a sequence of *r* codons, the total number of synonymous sites (*S*) and non-synonymous sites (*N*) is given by:

$$S = \sum_{j=1}^{r} s_j \text{ and } N = 3r - s,$$

where s_j is the value of *s* at the j^{th} codon.

To compare two homologous sequences, the averages of *S* and *N* for the two sequences are used. The number of synonymous differences (s_d) and the number of non-synonymous differences (n_d) between the sequences are calculated at each codon in turn. When more than one nucleotide difference exists between a pair of codons, there are multiple evolutionary paths between one codon and the other which may involve different numbers of synonymous or non-synonymous changes. Under the method of Nei and Gojobori, all such evolutionary pathways are enumerated and are assumed to have occurred with equal probability. The quantities s_d and n_d are then averaged over all evolutionary paths.

The total number of synonymous differences between the two sequences is given by:

$$S_d = \sum_{j=1}^{r} s_{dj} \text{ and } N_d = \sum_{j=1}^{r} n_{dj}.$$

The proportion of synonymous differences is then given by p_S=S_d/S and the proportion of non-synonymous differences is given by p_N=N_d/N, where S and N are the averaged values across the two sequences. The formula of Jukes and Cantor (1969) is used to estimate the number of synonymous or non-synonymous substitutions per site (d_S and d_N respectively) as follows (Felsenstein 1985b):

$$d_S = -\frac{3}{4}\ln\left(1-\frac{4}{3}p_S\right) \text{and } d_N = -\frac{3}{4}\ln\left(1-\frac{4}{3}p_N\right).$$

Nei and Gojobori (1986) simplified of the method of Miyata and Yasunaga (1980) by assigning equal weights to evolutionary paths rather than preferentially weighting for synonymous substitutions over non-synonymous substitutions. Nei and Gojobori (1986) showed that this simplifying assumption did not substantially increase error in d_N /d_S estimates, and that it made the method less computationally prohibitive to users. The method of Zhang *et al*. (1998) extended that of Nei and Gojobori (1986), by allowing for a higher rate of transitions than transversions when calculating the number of synonymous and non-synonymous sites.

Pairwise distance methods have been criticised for not incorporating phylogenetic relationships between sequences. Failing to account for shared ancestry can lead to overestimates of the number of inferred substitutions, which may result from a single ancestral substitution rather than multiple independent changes (Felsenstein 1985b; Kosakovsky Pond *et al.* 2009). Classes of methods which count the number of synonymous and non-synonymous substitutions along ancestral sequences based upon phylogenetic reconstructions have now been developed, for example the parsimony-based method of Suzuki and Gojobori (1999), and the single likelihood ancestor counting methods of Kosakovsky Pond and Frost (2005).

Maximum likelihood codon methods (Goldman and Yang 1994; Muse and Gaut 1994), which employ an explicit probabilistic model of codon evolution, have also been developed for providing gene-wide d_N/d_S estimates. Both the Goldman-Yang and Muse-Gaut methods model codon replacement as a Markov process, although the parameterisation differs between the models. The Muse-Gaut model is parameterised in terms of synonymous and non-synonymous substitution rates, whilst the Goldman-Yang model fixes the rate of synonymous evolution to 1 and estimates the non-synonymous rate relative to this. Under the ML framework, both methods can be used to formally test for an elevated level of non-synonymous change relative to the amount of synonymous change.

2.11.3 Methods for calculating site-to-site $d_\mathrm{N}/d_\mathrm{S}$ ratios

Whilst methods for detecting positive selection across a gene can provide an overall indicator of selective pressure, they are thought to be conservative in terms of detecting positive selection since only a small number of positively selected sites may exist in a region of overall functional constraint (Yang and Bielawski 2000). Phylogenetic approaches for detecting individual amino acid sites under positive selection were first described by Nielsen and Yang (1998). The method of Nielsen and Yang (1998) uses maximum-likelihood to calculate d_N and d_S along an alignment, with the ratio $d_\mathrm{N}/d_\mathrm{S}$ allowed to vary across sites. Likelihood ratio tests may be used to calculate the probability that a given site is a member of the class of positively selected sites, conditional upon information at the tips of the phylogeny for the sequences.

The method of Nielsen and Yang (1998) was later extended to allow different distributions of $d_\mathrm{N}/d_\mathrm{S}$ across amino acid sites (Yang *et al.* 2000). A further development, known as an 'empirical Bayes' approach, allowed prior beliefs about the nature of the selective pressure acting upon particular sites to be incorporated

(Yang and Swanson 2002). The development of a significance test for whether a particular site was under positive selection (Swanson *et al.* 2003) meant that individual amino acid sites with potential biological significance could be identified, rather than merely providing evidence for a class of positively selected sites. A Bayesian method which reports the posterior probability of each amino acid site being under positive selection, and accounts for uncertainty in model parameters such as branch lengths and substitution parameters, has also been developed (Huelsenbeck and Dyer 2004). This is known as a 'hierarchical Bayes' approach.

Kosakovsky Pond and Frost (2005) identified three classes of methods for estimating d_N and d_S to study selection on a site-by-site basis. 'Counting methods' involve reconstructing ancestral sequences and counting the number of synonymous and non-synonymous changes throughout the evolutionary history of each codon. For example, the stochastic mutational mapping method of Nielsen (2001, 2002) has been used to count numbers of synonymous and non-synonymous substitutions and detect positively selected sites (Nielsen and Huelsenbeck 2002). A counting method known as single likelihood ancestor counting (SLAC) method has been developed by Kosakovsky Pond and Frost (2005).

The second class of methods involves modelling synonymous and non-synonymous rates according to a predefined distribution, as implemented in the methods of Nielsen and Yang (1998), Yang et al. (2000) and Huelsenbeck and Dyer (2004). These methods are termed random effects likelihood (REL) methods. Finally, fixed effects likelihood (FEL) methods estimate d_N and d_S directly for each site. SLAC, REL, and FEL methods are all based upon an underlying phylogeny and a codon substitution model, and have been implemented in the HyPhy software. Kosakovsky Pond *et al*. (2005) found similar levels of type I and type II errors, and consistent estimates of d_N / d_S, across all three methods during simulation studies and tests on sequence data. A phylogenetic maximum likelihood test for accelerated substitution towards a particular amino acid (and the identity of that residue) at individual sites along a protein alignment (a method known as 'DEPS') has recently been developed (Kosakovsky Pond *et al.* 2008).

2.12 Viral phylogenies and effective population size

A Wright-Fisher population is an idealised population with non-overlapping generations where each individual contributes an infinite number of gametes to a gene pool and all individuals are replaced in each generation (Fisher 1930; Wright 1931). The effective size (N_e) of a population is defined to be the size of a Wright-Fisher population whose genetic diversity would change at the same rate as the census population size (see Rodrigo (2009)). This leads to a coalescent-based interpretation of N_e, known as the coalescent effective population size, which is the value of N_e which provides the same distribution of coalescence times as would be obtained for the actual biological population under consideration. The meaning of coalescent effective population size, and methods for its calculation, has been discussed in greater detail by Sjodin *et al*. (2005) and Wakeley and Sargsyan (2009).

Nee *et al*. (1995) and Kuhner, Felsenstein *et al*. (Kuhner *et al.* 1995; Beerli and Felsenstein 1999) independently recognised that the coalescent theory could be used to make inferences about the history of a population. Nee *et al*. (1995) achieved this by plotting the Number of Lineages in a phylogeny as a Function of Time (NLFT). The NLFT approach was later extended by Pybus, Rambaut and Harvey (2000) and Strimmer and Pybus (2001), who used the notion that the effective size of a population could change at coalescent events to generate plots of N_e over time, known as skyline plots, in a maximum likelihood framework. These methods have been used to investigate the demographic history of a population directly from a sample of genetic sequences, without the requirement for a pre-specified demographic model. Whereas the classic skyline plot of Pybus *et al*. (2000) forced a change in N_e at each coalescent event, the generalised skyline of Strimmer and Pybus (2001) provided smoothing by allowing contiguous coalescent events to be grouped together across a period of constant effective population size.

Skyline demographic inference in a Bayesian framework was introduced by Drummond *et al*. (2005) and Opgen-Rhein *et al*. (2005), who used MCMC to obtain posterior samples of N_e through time. Unlike ML approaches, Bayesian methods can incorporate uncertainty in the phylogenetic and coalescent process. One disadvantage of the methods of Drummond *et al*. (2005) and Opgen-Rhein *et al*. (2005) is the requirement to inform the number of change-points in effective population size over time *a priori*. A Gaussian Markov random field (gmrf) prior with time-aware smoothing (penalising changes in N_e over short time intervals compared to over longer time intervals) was introduced by Minin *et al*. (2008) to resolve this problem. This 'Bayesian skyride' method does not require strong prior decisions about the number of change-points in N_e, and still captures important changes in N_e over time (Minin *et al.* 2008). See Chapter 7 (Section 7.4.5) for further discussion of coalescent-based inference of N_e, and how the coalescence rate can be related to epidemiological parameters such as incidence and prevalence.

2.13 Detecting recombination

Recombination can confound attempts to reconstruct phylogenies since recombinant sequences will cluster in different positions in phylogenies constructed on either side of a recombination breakpoint (Posada and Crandall 2002). Failure to account for recombination can lead to an abundance of false positives when identifying positively selected sites, although gene-wide estimates are less affected (Anisimova *et al.* 2003; Shriner *et al.* 2003). Estimates of divergence dates may also be affected by recombination (Schierup and Hein 2000; Worobey 2001; Schierup and Forsberg 2003), or at least the variance of such estimates may be increased (Lemey *et al.* 2004). Although it is widely held that recombination between homologous influenza segments does not occur naturally in influenza viruses (e.g. Boni *et al*. (2008)), its presence was screened for in all of the datasets analysed in this thesis.

Several methods for detecting recombination from sequence alignments have been developed and their effectiveness has been compared via simulation studies (Posada and Crandall 2001). A genetic algorithm method for recombination detection (GARD) has been developed, which explicitly models site-to-site rate variation and thus has a lower rate of false-positive breakpoints than other methods (Kosakovsky Pond *et al.* 2006a). The special case where only one breakpoint is allowed in the alignment is known as a single breakpoint analysis (SBP), and is particularly useful where rapid screening is required to confirm that recombination is not present in a dataset. Both GARD and SBP are implemented in the HyPhy software.

Under the SBP algorithm, an initial Neighbor-Joining tree is constructed for the entire alignment and the small-sample AIC score[1] (AICc) (Sugiura 1978) is computed, using maximum likelihood to calculate substitution rate parameters and branch lengths. Substitution parameter estimates are then fixed for all subsequent inferences. For all possible recombination breakpoints, neighbour-joining trees are constructed separately for the portions of the alignment on each side of the potential breakpoint. The modified AICc score for the model fitting branch lengths to the portions of the alignment on either side of the breakpoint separately (the partitioned model) is compared with that for the whole alignment (the un-partitioned model). A lower AICc score for the partitioned model than for the un-partitioned model provides evidence for recombination in the dataset, and the relative support for different positions for the breakpoint can be evaluated via Akaike weights (Akaike 1983). The Shimodaira and Hasegawa test (Shimodaira and Hasegawa 1999) is used to assess whether perceived differences in the trees are simply due to variation in branch lengths or if there is genuine topological incongruence.

[1] To obtain an AICc score, the penalty term of twice the number of free parameters from the usual AIC is multiplied by $N/(N-p-1)$, where N is the number of columns in the alignment and p is the number of free parameters. The use of AICc has been advocated when the number of alignment columns is less than 40 times the number of parameters in the model (Burnham and Anderson 2003). The number of parameters in the model is the sum of the number of branches in the trees fitted to either side of the potential breakpoint and the number of parameters in the evolutionary model.

The ‘tree-order scan’ method for visualising phylogenetic relationships between taxa along the length of an alignment, which can be used to identify phylogenetic discordance, was developed by Simmonds and Midgley (2005) and is available in the SSE software (Simmonds 2012). Bootstrapped, rooted neighbour-joining trees are constructed from overlapping sequence fragments using a sliding window approach (with the length of the fragment on which to construct the phylogeny, and the increment by which to shift the window, chosen by the user). Changes in the phylogenetic position of sequences and/or clades with greater than a specified level of bootstrap support between consecutive fragments are reported. A plot is then constructed to display the position of the sequences in the phylogeny against the genome position. If the phylogenetic relationships did not change across the genome, the plot would contain a horizontal line for each sequence with no crossing over. Phylogenetic discordance is detected as crossing over of lines and is indicative of recombination.

2.14 Viral sequence data

All of the influenza data analysed in this thesis were obtained from the NCBI Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) (Bao *et al.* 2008), a database of influenza genome sequences from GenBank and the NIAID Influenza Genome Sequencing project. The database contained approximately 2,500 avian full genome sequences and over 6000 full-length protein coding avian HA sequences as at February 2011. Sequences in the database are generally labelled by attributes such as virus subtype, location and host species of isolation and the date of sampling.

Chapter 3

Phylogenetic analysis of H7 avian influenza virus HA

3 Phylogenetic analysis of H7 avian influenza virus HA

3.1 Chapter Summary

The only recorded highly pathogenic avian influenza viruses to date have been of HA subtype H5 or H7. Avian influenza viruses of subtype H7 circulate globally in wild birds, cause disruption to the poultry industry and have the potential to cause infection in humans. In this chapter, I conduct a molecular evolutionary analysis of all H7 avian HA sequences available on the NCBI database. Maximum likelihood and time-scaled Bayesian phylogenetic analyses are performed to examine global evolutionary relationships. Consistent with earlier studies, I demonstrate clustering of H7 avian influenza HA sequences into major geographical lineages (Eurasian and American). I also provide evidence for the existence of distinct cleavage site motifs for the Eurasian and American lineages. The phylogenies provide evidence for the repeated transfer of influenza virus from wild to domesticated birds, multiple independent emergences of highly pathogenic virus and frequent reassortment of different NA subtypes onto the H7 HA background.

3.2 Chapter Aims

- Investigate phylogeographic relationships between all available H7 HA avian influenza sequences in the NCBI database
- Investigate the distribution of different NA subtype backgrounds across the H7 HA phylogenies, as well as the distribution of highly pathogenic sequences and sequences from wild or domestic birds
- Summarise H7 HA cleavage site motifs – a marker for virus pathogenicity
- Identify suitable subsets of the H7 sequence data for further evolutionary analysis

3.3 Introduction

Low pathogenic (LP) avian influenza viruses have been documented for all haemagglutinin (HA) subtypes; however, to date the only known highly pathogenic (HP) viruses have been of the H5 or H7 subtype. H7 avian influenza viruses have been associated with poultry outbreaks for many decades, with the earliest highly pathogenic fowl plague virus (FPV) sequences, of subtypes H7N1 and H7N7, originating from Europe between 1927 and 1945. Outbreaks of HP H7 avian influenza in poultry have occurred in North and South America, Europe, Asia and Australia, causing substantial economic loss and disruption. For example, the H7N1 avian influenza outbreak in Northern Italy in 1999-2000, which was caused by a circulating LPAI virus mutating into a highly pathogenic form, resulted in the loss of 16 million birds. H7 avian influenza viruses of both low and high pathogenicity have been found to cause infection in humans (Kurtz *et al.* 1996; Fouchier *et al.* 2004), with one case of human fatality having been reported (Fouchier *et al.* 2004).

The evolution and continued circulation of H5 avian influenza, in particular the antigenic HA segment, has been studied extensively due to concern over mortality in poultry, wild birds and humans infected with HP H5N1 (e.g. Li *et al*. (2004), Vijaykrishna *et al*. (2008a) and Vijaykrishna *et al*. (2008b)). In contrast, at the time that this analysis was carried out, comprehensive and up-to-date phylogenetic analyses of H7 avian influenza were lacking, with the largest and most recent being a study of 53 HA1 sequences isolated between 1927 and 1999 (Banks *et al.* 2000). Recent evidence that H7N2 viruses circulating in North American poultry between 1994 and 2006 acquired human-like receptor binding affinities (Belser *et al.* 2008) suggests that contemporary H7 viruses may be an epidemiological threat to humans, as well as an economic burden. It is therefore important to reveal global evolutionary relationships between H7 avian influenza viruses, for example with respect to geographical location, host species or virus pathogenicity. Such findings could help to target future surveillance studies, facilitating the early detection of a potential pandemic strain.

Routine surveillance of live bird markets in both North America and Asia, intensive sampling of viruses from influenza outbreaks in poultry and increased sampling of influenza viruses from wild birds means that information can now be combined to give a global picture of how avian influenza viruses are circulating, in addition to looking at their evolution in different demographic scenarios. In this chapter, I perform a phylogenetic analysis of all unique full-length avian influenza H7 HA sequences available in the NCBI database. In addition to maximum likelihood analysis, I use the BEAST software to construct time-scaled phylogenies by incorporating the date of sampling in the analysis. Under the Bayesian methodology employed by BEAST, uncertainty in the phylogenetic process is accounted for by using MCMC to obtain a sample of trees and parameters from the posterior distribution. I also attempt to resolve the phylogenetic position of the early European fowl plague virus isolates, which have puzzlingly been found to group with contemporary Australian sequences in previous phylogenetic analyses. I examine the phylogenetic clustering of available H7 avian influenza virus HA sequences, in terms of NA subtype of the virus, as well as the geographical location and avian host from which the virus was isolated.

3.4 Methods

3.4.1 Dataset

All available unique, full-length avian influenza HA protein-coding sequences of the H7 subtype were downloaded from the NCBI database. In cases where the same virus (identified by the NBCI virus name) had been sequenced more than once, only one such sequence (the longest) was retained. Sequences were aligned manually using BioEdit (Hall 1999). The HA1/HA2 cleavage region (amino acids between the motif P**P at the 3′ end of the HA1 coding region - where ‘P’ is the standard abbreviation from proline and ‘*’ represents any other amino acid residue - and the motif GLF at the start of the HA2 coding region) was removed for the phylogenetic analysis. The length of the alignment to be analysed was 1671 nucleotides.

Sequences were labelled according to the NA subtype of the virus, their avian host (where known, the species and whether the bird was wild or domestic), the geographical location from which the bird was sampled and the year of sampling. ‘Domestic’ birds included farmed, caged and aviary birds. Where possible, classification of the sequences into highly pathogenic (HP) or low pathogenic (LP) was made by searching the literature for studies confirming the pathogenic status of the virus using laboratory testing. Where no record of the pathogenicity of an isolate could be found, sequences were classified as HP if they possessed a motif at the HA1/HA2 cleavage region which was the same as that of a previously confirmed HP virus, in accordance with the guidelines of Alexander (2000). Sequences with a novel cleavage site motif which had not been previously documented as either HP or LP were not labelled by pathogenicity.

3.4.2 Phylogenetic Analysis

Phylogenetic analysis of the avian H7 HA sequences was performed using neighbor-joining, maximum likelihood and Bayesian methodologies. Neighbor-joining (NJ) trees were constructed using MEGA version 5.05, under the TN93 (Tamura and Nei 1993) nucleotide substitution model, with gamma distributed rate heterogeneity across sites and allowing for heterogeneity across lineages. Maximum likelihood (ML) phylogenies were constructed in PhyML version 3.0 (Guindon *et al.* 2010), with a general time reversible (GTR) model of nucleotide substitution (Tavaré 1986), gamma distributed rate heterogeneity across sites and four rate categories. Bootstrapped replicates of phylogenies were sampled to assess support for clades.

Bayesian phylogenetic inference was performed using the BEAST software (Drummond and Rambaut 2007) and the year of sampling of the sequences was used to obtain trees with an explicit timescale. Preliminary analysis using strict and relaxed uncorrelated lognormal (ucln) molecular clocks suggested that a relaxed clock model provided the best fit to the data on the basis of Bayes factor testing, as has previously been found to be the case in evolutionary studies of avian influenza, for example by Vijaykrishna *et al*. (2008a). Analyses were carried out under a relaxed demographic model: the gmrf Bayesian skyride with time-aware smoothing (Minin and Suchard 2008a).

Posterior estimates of evolutionary parameters and phylogenies were obtained from sampling at intervals of at least 10,000 generations over a period of at least 100 million generations for all BEAST runs. For each combination of settings, convergence of the MCMC chain was assessed for multiple independent runs by manually inspecting the chain traces in the Tracer software. The effective sample size (ESS) was greater than 200 for all parameters in all runs from which trees and evolutionary parameters are reported. After removing the first 10% of samples for burnin, 9,000 posterior samples were available for each run. The FigTree software was used to visualise summary phylogenies (maximum clade credibility trees from BEAST, or bootstrapped NJ or ML trees). The tips of the trees were coloured or labelled by various factors, including the NA subtype of the virus and whether the viruses were HP or LP, or from wild or domestic avian hosts.

Initial BEAST analyses were performed upon a subsampled version (206 sequences) of the full avian H7 HA dataset, in an attempt to assess clustering into major geographical lineages whilst overcoming computational constraints. Sequences were subsampled to preserve diversity with respect to location, host, pathogenicity, NA subtype background and year of sampling. Analyses were performed both including and excluding the FPV sequences. In addition, separate North American and Eurasian/African datasets were created for individual analysis with BEAST due to the high level of divergence between the geographical lineages identified in the NJ and ML analyses. Subsampling of avian H7 HA sequences from the NCBI database was not performed upon the individual North American and Eurasian datasets. Only sequences for which the pathogenicity, host origin and NA subtype of the virus were known were included in the BEAST analysis.

The Path-O-Gen software (http://tree.bio.ed.ac.uk/software/pathogen/) was used to investigate the temporal signal in the neighbour-joining and maximum likelihood trees, which had not been inferred under the assumption of a molecular clock. Root-to-tip divergence within individual geographical clades (North American, Eurasian and Australian) was investigated using individual NJ trees constructed for each geographical clade in order to obtain preliminary estimates of the rate of nucleotide substitution, to assess how clock-like the data was and to identify any outlying taxa. An outgroup was chosen from a different geographical clade for each dataset (e.g. a North American sequence for the Eurasian dataset) and the best fitting root of the tree was selected from which to calculate root-to-tip distances in Path-O-Gen. In addition, the placing of the Eurasian fowl plague viruses relative to later European sequences was examined using a plot of the root-to-tip distance against the date of sampling, in order to consider whether their position in the tree was as would be expected under clock-like molecular evolution.

3.5 Results

3.5.1 Distribution of H7 avian influenza HA sequences

After removing identical sequences, multiple sequencings of the same strain and sequences from viruses of unknown NA subtype, a total of 470 full-length avian H7 HA protein coding sequences were available on the NCBI Influenza Resource. Of these, 295 were from wild birds and 155 were from domestic birds, whilst no information as to whether the host was a wild or domestic bird was available for the remaining 19 sequences (labelled ‘duck’ or ‘softbill’ with no further reference to the host status available in the literature). Domestic birds included farmed birds (e.g. chickens, farmed ducks and ostrich) as well as caged and aviary birds such as parakeets and macaws. The taxonomic orders of the avian hosts from which the H7 influenza viruses were isolated was recorded (Table 3.1 and Figure 3.1).

Taxonomic Order	Examples in avian H7 HA dataset
Anseriformes (Ans.)	Duck, goose, swan, teal, widgeon, northern shoveler, garganey, northern pintail
Galliformes (Gal.)	Turkey, grouse, chicken, quail, pheasant, guinea-fowl, chukar
Struthioniformes (Str.)	Ostrich, emu
Passeriformes (Pas.)	Starling, fairy bluebird, common iora, pekin robin, magpie
Psittaciformes (Psi.)	Parrot, parakeet, conure, macaw
Charadriiformes (Cha.)	Gull, ruddy turnstone, shorebird/wader, red knot, tern, sanderling
Rheiformes (Rhe.)	Rhea

Table 3.1
Classification of birds by taxonomic order. The names of all avian hosts for which sequence information was available were extracted from the H7 HA dataset. The left-hand column contains the avian host order and its abbreviation, whilst the right-hand column lists the birds of that order for which an influenza HA sequence sampled from that host was present in the dataset.

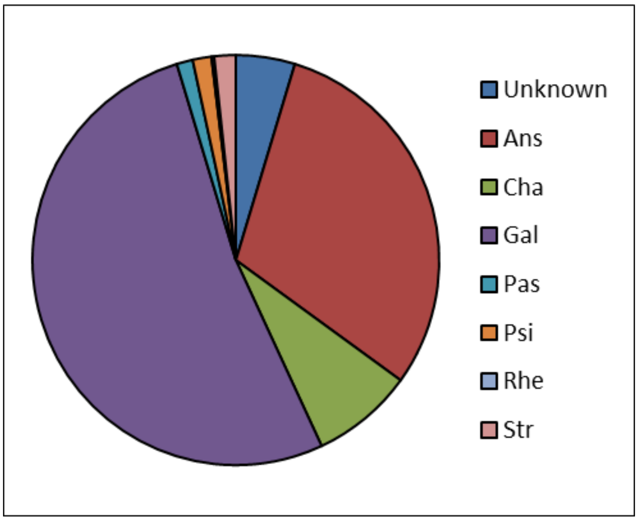


Figure 3.1
Avian H7 HA influenza sequences by taxonomic order of host. The number of sequences available in the NCBI influenza virus database, from avian hosts of different orders, is presented. Some host orders represent only wild (e.g. Charadriiformes - Cha.) or domestic hosts (e.g. Galliformes - Gal.), whilst others (e.g. Anseriformes - Ans.) include both wild and domestic birds. The full names and abbreviations used for taxonomic orders in the key are listed in Table 3.1.

22 sequences were from avian hosts whose order could not be determined from the sequence label or a search of the literature, and included sequences labelled ‘softbill’, ‘non-psittacine’, ‘fowl’, ‘avian’ and ‘wild bird faeces’. The most common bird order from which H7 HA sequences had been isolated (142 of the 470 samples) was Galliformes, which are terrestrial poultry such as chickens and turkeys. The second most common avian host order corresponding to the avian influenza H7 HA sequences was Anseriformes, which includes ducks and geese. 106 of the sequences from Anseriformes were from wild birds and 40 were from domestic birds; it was not known whether the host was wild or domestic in 17 cases. Other avian host orders from which viruses were sequenced included Charadriiformes (gulls and terns: 38 sequences, all from wild birds), Passeriformes (8 sequences from wild birds), Psittaciformes (7 sequences from caged birds), Struthiformes (8 sequences from farmed ostriches and emus) and Rheiformes (1 sequence from a farmed rhea).

The earliest H7 HA sequence available was sampled from a domestic bird in 1927, and isolates from domestic birds were present from each decade thereafter until the present day, except for the 1950s. However, only a small number of sequences were available prior to 1990 (Figure 3.2). The earliest avian H7 HA sequences from wild birds date from the 1970s, although just 19 pre-2000 sequences are available. The large number of recent wild bird sequences (136 since 2000) is a result of increased surveillance for avian influenza, particularly following the emergence of highly pathogenic H5N1 avian influenza viruses.

H7 avian influenza HA sequences from wild birds were available in the NCBI database for viruses of all NA subtypes (N1 – N9), although the subtypes differed greatly in abundance (Figure 3.3). All but two H7 serotypes (H7N5 and H7N8 – which were only present once and four times respectively amongst the wild bird sequences) were present in domestic birds. H7N3 and H7N7 were the most frequently occurring serotypes amongst avian H7 HA sequences from wild birds (70 and 51 sequences respectively), with less than 10 sequences present for all other NA subtype backgrounds. H7N2 was the most frequently sampled subtype amongst domestic birds (131 sequences), followed by N3, N1 and N7, with less than 10 sequences available for each of the remaining subtypes.

.

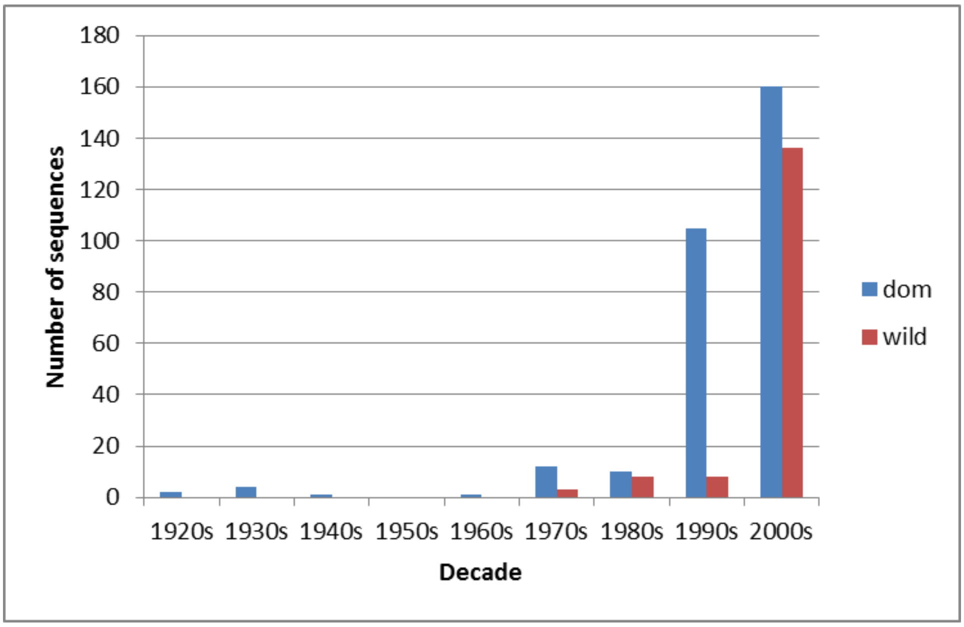


Figure 3.2
Avian H7 HA influenza sequences by decade. The number of sequences from wild and domestic (dom) birds available in the NCBI influenza virus database is shown for each decade from the 1920s (corresponding to the earliest available isolates) to the present day.

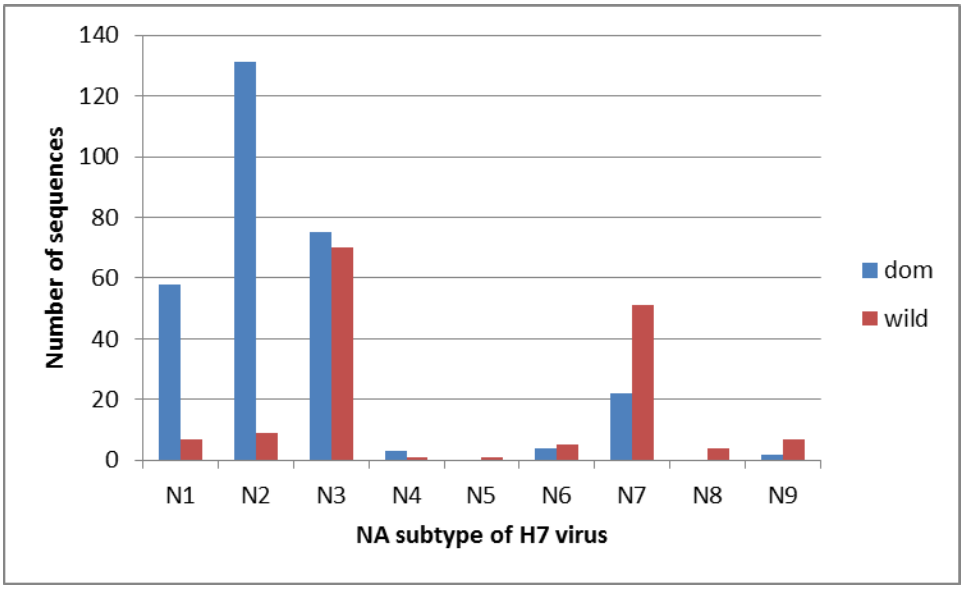


Figure 3.3
Avian H7 HA influenza sequences by NA subtype. The number of sequences from wild and domestic (dom) birds available in the NCBI influenza virus database is shown for each NA subtype, N1-N9.

The continent of origin was also recorded for avian H7 HA sequences (Figure 3.4). 51% of the available avian H7 HA sequences were from birds in North America (encompassing two sequences isolated from wild birds in Guatemala in Central America). H7 viruses were found on all NA subtype backgrounds in North America. Notable groups of sequences from domestic birds in North America include low pathogenic H7N2 isolates sampled from live bird markets between 1994 and 2006, which were all of low pathogenicity. The only HP isolates from North America corresponded to outbreaks of H7N3 in Canada: in British Columbia in 2004 and Saskatchewan in 2007. From 2006 onwards, large numbers of sequences were obtained from wild birds of orders Anseriformes and Charadriiformes along the East and West coasts of America, with H7N3 being the most common subtype. Just 2% of unique avian H7 HA sequences in the NCBI database were from South American birds. These were from an outbreak of highly pathogenic H7N3 in domestic birds in Chile in 2002, and an H7N3 virus isolated from a wild duck in Bolivia in 2001.

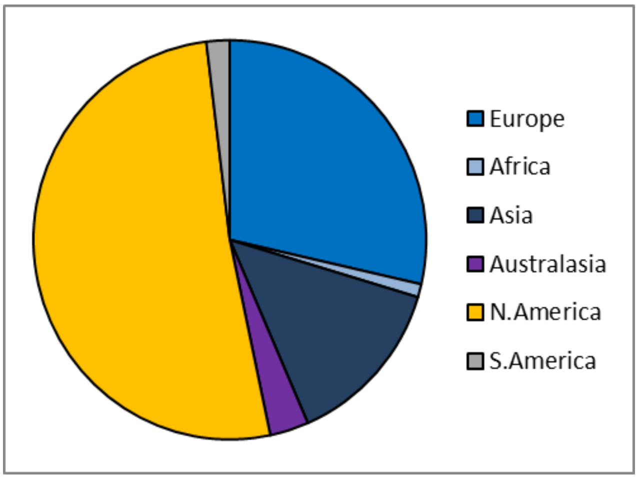


Figure 3.4
Avian H7 HA influenza sequences by geographical region. The number of sequences available in the NCBI influenza virus database, from major geographical regions, is presented. Under the conventions for influenza virus sequence labelling, the location refers to the geographical region where the sample was taken from the bird (except for some of the early fowl plague virus isolates, e.g. A/fowl/Dobson/1927 and A/fowl/Weybridge, which appear to have been labelled by the location in which the sequencing took place).

28% of the avian H7 HA sequences in the NCBI database were sampled in Europe. Representative sequences were present for many outbreaks of avian influenza in domestic birds. These outbreaks included highly pathogenic H7N7 isolated from geese and chickens in Germany in 1979, and sequences from H7N1 viruses of both low and high pathogenicity sampled during an avian influenza outbreak in Italy between 1999 and 2000. Low pathogenic H7N3 sequences were also sampled from galliform birds in Italy in 2002. Three sequences were available from the outbreak of highly pathogenic H7N7 which occurred on poultry farms in the Netherlands and Germany in 2003, leading to the death of a veterinarian (Fouchier *et al.* 2004). Fourteen low pathogenic H7N7 sequences from mallards in Sweden in 2002 were present; however, wild birds in Europe appear to have been sampled sporadically overall, in contrast to large number of sequences available from surveillance of wild birds in North America.

After North America and Europe, the most common continent of origin for the avian H7 HA sequences was Asia, from which 14% of the sequences were sampled. HP and LP H7N3 sequences were available from chickens in Pakistan between 1995 and 2004, and LP H7N6 sequences isolated on quail farms in Japan in 2006 were also present. In addition, LP sequences from wild anseriform birds in Japan and Korea were sampled from 2006 onwards, and the predominant subtype amongst these was H7N7 (in contrast to H7N3 in North America). A small number of H7 HA sequences from Africa were also present in the NCBI database. These were isolated from farmed ostriches in the 1990s, were low pathogenic and of subtypes H7N1 and H7N7.

Available H7 HA sequences from Australasia were sampled in Australia, New Zealand and Tasmania and were from viruses with NA background subtypes N2, N3, N4, N6 and N7. The sequences were sampled between 1976 and 2007. 3 out of the 14 sequences were low pathogenic and were isolated from ducks, whilst 11 were highly pathogenic and were isolated from farmed chickens and emus (9 out of the 11 HP sequences) or wild starlings (2 out of the 11 HP sequences). It has previously been reported that the H7N7 viruses isolated from starlings in 1985 were closely related to HP chicken viruses of the same subtype, also isolated in Victoria in 1985 (Nestorowicz *et al.* 1987).

3.5.2 Molecular analysis of the H7 HA cleavage site

43 different motifs were observed at the HA cleavage site (Appendix A, Table A1). 21 of these motifs corresponded to LP viruses, 20 to HP viruses, and 2 to viruses whose pathogenicity could not be determined from the literature. Two motifs, PEIPKGR (observed 131 times) and PENPKTR (observed 130 times), were particularly dominant amongst the LP sequences and corresponded to Eurasian/African and North/South American sequences respectively. These motifs were present in isolates from wild and domestic birds from the early 1970s until 2009 (the latest sequences available), and were obtained from H7 avian influenza viruses on most NA subtype backgrounds (7 out of the 9 for the Eurasian motif, and 8 out of 9 for the American motif). The longstanding predominance of these motifs, without any crossover between the major geographical regions, is another indicator of the lack of transatlantic exchange of avian influenza viruses. Excluding the softbill isolates from California, which may have been imported from Europe (see Section 3.5.4), there were no HA cleavage site motifs which were found in both Eurasian and American sequences.

None of the LP isolates had amino acid insertions at the HA cleavage region, although motifs where replacement by basic amino acids (histidine (H), lysine (K) or arginine (R)) had taken place at the cleavage site were observed amongst the LP North American H7N2 isolates (see Section 4.6.4 for further discussion of North American H7N2). All HP sequences possessed at least two basic amino acid insertions at the HA cleavage site, and many had basic amino acid substitutions at the cleavage site such that all except one HP motif (from the Italian H7N1 outbreak of 1999-2000) contained at least 5 basic amino acids. 6 out of the 20 HP motifs also had insertions of non-basic amino acids at the cleavage site region.

3.5.3 Phylogenetic position of early European fowl plague virus sequences

7 highly pathogenic H7 HA FPV sequences were present in the NCBI database. These were sampled between 1927 and 1945 and were of subtypes H7N1 and H7N7. All except one (A/fowl/Egypt/1945(H7N1)) were sampled in Europe. It was not always possible to find information regarding the country of sampling or the host, although all sequences were obtained from domestic birds. All FPV isolates had six basic amino acids (in the motif KKRRKR or KKRKKR) at the HA cleavage site.

The European fowl plague viruses sampled between 1927 and 1945 clustered together with bootstrap values of greater or equal to 95% in the neighbour-joining and ML analyses. The FPV sequences clustered most closely with the Australian sequences in the NJ tree (Appendix A, Figure A1), consistent with an earlier maximum likelihood analysis (Banks *et al.* 2000) and a recent study (Bulach *et al.* 2010) which used the minimum evolution method to construct phylogenies. However, in the ML tree (Figure 3.5), the FPV sequences fell basal to the Eurasian and African sequences (bootstrap value = 51%), with the Australian sequences basal to both the FPV and Eurasian/African sequences (bootstrap value = 100%). The low bootstrap value for the clustering of the FPV sequences with the Eurasian and African sequences, along with different clustering behaviour found by previous studies and the NJ analysis, suggest uncertainty regarding the phylogenetic position of the FPV sequences.

It may be observed from the root-to-tip divergence plot for an NJ tree constructed from FPV and later Eurasian sequences (Appendix A, Figure A2) that the FPV sequences are further from the root of the tree than would be expected under a clock-like model of evolution. In fact, the FPV sequences are positioned a similar distance from the root to the Eurasian H7 HA sequences sampled more than 50 years later. One cannot exclude the possibility that the 1927-1945 sequences have been affected by extensive serial passage in eggs which was necessary as storage at -70°C was not available until the 1950s; such extensive passage could distort estimates of substitution rates or detection of positive selection. The FPV sequences were therefore excluded from subsequent analyses presented in this thesis and had previously been omitted from other studies of avian influenza evolution (e.g. Chen and Holmes (2006)) on the basis that they are highly cultured.

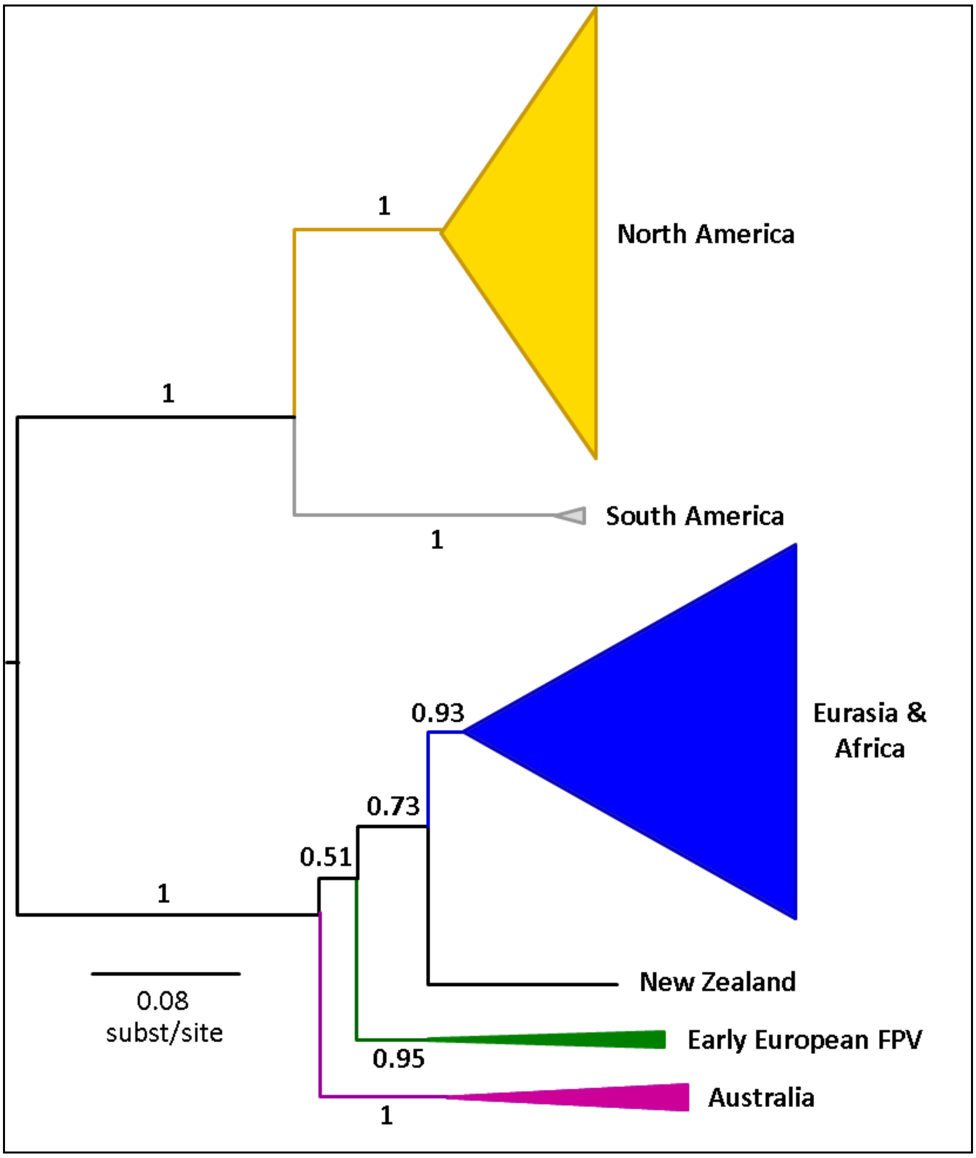


Figure 3.5
Maximum-likelihood phylogeny of Avian H7 HA influenza sequences. The phylogeny was constructed in PHYML, using a GTR model of nucleotide substitution and gamma distributed rate heterogeneity across sites, with six rate categories. The tree was rooted to an H15 HA outgroup sequence (removed from figure for visualisation purposes). Clades corresponding to major geographical lineages were identified and collapsed so that their sizes were proportional to the number of sequences at the tips of the tree in each clade. 100 bootstrap replicates were performed and bootstrap support values are reported as the proportion of the bootstrap replicates in which those sequences clustered together.

3.5.4 Clustering of avian influenza H7 HA into major geographical lineages

Phylogenetic analysis of avian H7 HA sequences provided strong evidence of clustering into major geographical lineages. In the ML phylogeny (Figure 3.5), it is possible to observe a split into two major clades: one corresponding to sequences from North and South America, with the other corresponding to sequences from Europe, Africa, Asia and Australasia, consistent with an earlier report for H7 HA (Banks *et al.* 2000) and a general pattern observed for avian influenza (e.g. Olsen *et al*. (2006)). Both of these major geographical clades have a bootstrap support value of 100% in the Neighbor-joining and maximum likelihood trees.

The American sequences cluster into separate North and South American sub-clades in the NJ, ML and BEAST trees (each having an NJ and ML bootstrap value of 100%, or BEAST posterior probability of 1). A phylogenetic split into separate North and South American lineages had been previously suggested by Spackman *et al*. (2006). The two Central American wild bird sequences, A/blue-winged teal/Guatemala/CIP049-01/2008(H7N9) and A/blue-winged teal/Guatemala/CIP049-02/2008(H7N9), clustered with wild bird isolates in the North American clade, with their HA sequences being most closely related to a Californian isolate, A/northern shoveler/California/HKWF1026/2007(H7N3), sampled in the previous year (98% nucleotide sequence identity from BLAST similarity search).

Within the major clade consisting of Eurasian, African, Asian and Australian sequences, the Eurasian and African sequences (excluding the early FPV sequences) cluster together in the phylogenies with high support. As recently suggested by Bulach *et al*. (2010), the Australian sequences form their own clade (NJ bootstrap = 100%; ML bootstrap = 100%; BEAST posterior probability = 1). The Australian clade includes the Tasmanian sequence A/duck/Tasmania/277/2007(H7N2). However, the New Zealand sequence A/mallard/New Zealand/1365-355/2005(H7N7) does not cluster with the Australian sequences, falling immediately basal to the Eurasian/African clade in the NJ and ML analyses, although with low support in the NJ analysis (NJ bootstrap = 59%; ML bootstrap = 73%).

Individual root-to-tip divergence analyses of the North American and Australian datasets (Figure 3.6a and Figure 3.6b respectively) indicate a tendency for the distance between the root and the tips of the tree to increase with year of sampling. In the Australian dataset in particular, a strong temporal signal could be observed, with a Pearson product moment correlation coefficient of r = 0.896 in the linear regression of root-to-tip distance on sample date. In the Eurasian dataset (Figure 3.6c), tips corresponding to sequences sampled before 1990 appeared further from the root of the tree than would be expected under a strict molecular clock.

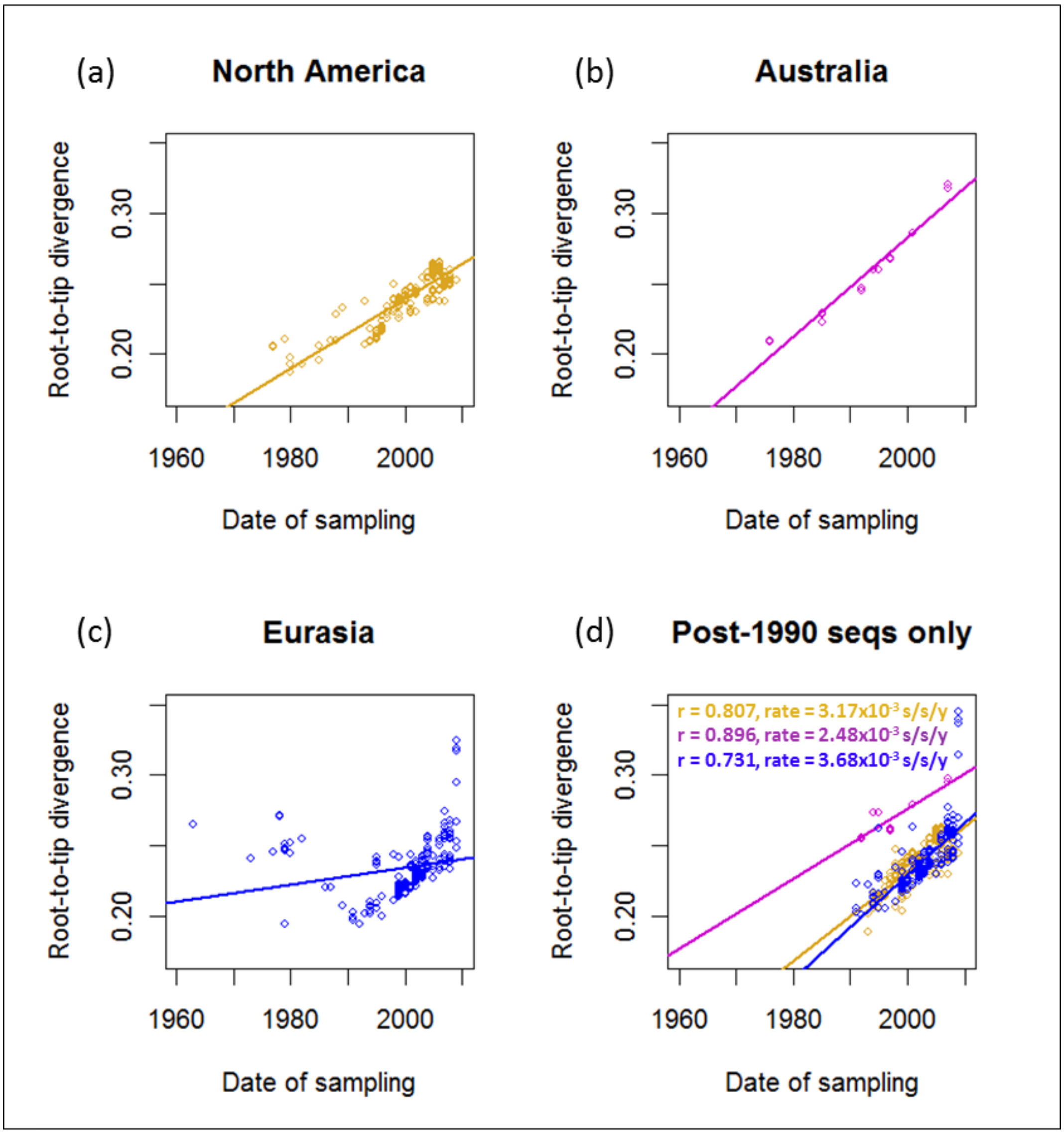


Figure 3.6
Root-to-tip distance plots for NJ phylogenies of Avian H7 HA influenza sequences from different regions. For each tree tip, the distance between that tip and the root of the tree was calculated using Path-O-Gen and plotted against the year of sampling. Plots (a),

(b) and (c) show the root-to-tip distances from NJ trees constructed from individual North American, Australian and Eurasian datasets respectively. Plot (d) shows root-to-tip distances for NJ trees constructed using only sequences sampled in, or after, 1990. Points are coloured corresponding to membership of major geographical clades: orange = North America; blue = Eurasia and Africa; purple = Australia. The Pearson product moment correlation coefficients are reported for each of the post-1990 analyses, as are the rates of nucleotide substitution calculated from the slope of the regression line of root-to-tip distance on date of sampling.

Initial estimates of the rate of nucleotide substitution could be obtained from the slope of the regression line of root-to-tip distance on year of sampling. Substitution rates were estimated from NJ trees constructed only from sequences sampled in, or after, 1990, following the root-to-tip analysis. Estimates of the rate of nucleotide substitution were similar for the North American, Eurasian and Australian datasets, as can be observed from the slopes of the regression lines in Figure 3.6d. These rates, of $3.17\text{x}10^{-3}$, $3.68\text{x}10^{-3}$ and $2.48\text{x}10^{-3}$ substitutions/site/year respectively, are consistent with previous estimates of the rate of avian influenza HA substitution such as those of Chen and Holmes (2006). Note that the lack of temporal diversity amongst the South American sequences (data were only from two time-points: 2001 and 2002) meant that it was not appropriate to perform a root-to-tip analysis for this continent.

Although all other avian H7 HA sequences clustered within the major geographical clade (North/South America or Eurasia/Africa/Australasia) corresponding to their location of sampling, three sequences sampled in North America fell within the Eurasian clade in the NJ and ML analyses (NJ bootstrap = 96%; ML bootstrap = 100%) and were not included in the BEAST analysis. Two of these sequences, A/softbill/California/33445-136/1992(H7N1) and A/softbill/CA/33445-158/1992(H7N1), clustered most closely with A/non-psittacine/England-Q/1985/89(H7N7), obtained from a quarantined bird in England three years earlier. This relationship has not been reported by other phylogenetic studies, perhaps because the Californian sequences were not submitted to GenBank until 2009. Since softbill birds such as the Pekin robin are Eurasian in origin and are frequently caged and transported for use as pets or in aviaries (Vince 1996), it is likely that

A/softbill/California was of Eurasian origin and entered North America by these means.

The other H7 avian influenza sequence which was sampled in North America, but which fell within the Eurasian and African clade, was A/Pekin robin/California/ 30412/1994(H7N1). This sequence clustered, with high support, with contemporaneous Eurasian H7N1 sequences from caged birds. A search of the literature confirmed that this bird was part of a shipment of birds into the United States, which were denied entry upon arrival) (Panigrahy and Senne 2003), again implicating the transportation of caged birds as a potential mechanism for the global transmission of avian influenza viruses.

3.5.5 Time-scaled phylogenetic analysis

The detailed BEAST summary phylogenies for the Eurasian/African and North American datasets (Figure 3.7 and Figure 3.8 respectively) show a strong tendency for H7 HA sequences from the same avian influenza outbreak to cluster together. Clustering of sequences from the same avian influenza outbreak was also observed in the NJ and ML trees. Notable outbreaks of H7 avian influenza for which clusters of sequences may be observed in the Eurasian tree include the 1999-2000 Italian H7N1 outbreak (posterior probability for clade = 1) and the 2002-2004 Italian H7N3 outbreak (posterior probability for clade = 0.996), both of which took place in domestic poultry (predominantly turkeys). The HA clades corresponding to the Italian H7N1 and H7N3 outbreaks fall in different parts of the tree. The H7N3 Italian outbreak sequences cluster with Italian H7N3 sequences sampled from wild birds in 2001. This indicates that the H7N3 outbreak arose through a separate introduction of virus from wild to domestic birds as suggested by Campitelli *et al.* (2004), rather than through reassortment of the HA segment from the Italian H7N1 outbreak onto an N3 genetic background. Previous analysis has also shown the Italian H7N1 HA sequences to be most closely related to an HA sequence from a wild bird (Banks *et al.* 2000).

The H7N1 Italian HA clade contained both HP and LP viruses. The HP viruses formed a monophyletic clade (posterior probability = 1) within the clade of Italian H7N1 sequences. In contrast, for H7N3 isolated from domestic chickens in Pakistan between 1995 and 2004, two distinct lineages can be observed which fall in different parts of the tree and correspond to HP and LP viruses. A previous analysis suggested that this was due to the maintenance of separate populations of HP and LP virus in Pakistan (Abbas *et al.* 2010). However, preliminary NJ, ML and BEAST analysis revealed that the LP isolates, A/chicken/Chakwal/NARC-35/2001(H7N3) and A/chicken/Pakistan/34668/1995 (H7N3) clustered most closely with an earlier sequence, A/parrot/Northern Ireland/VF7367/1973(H7N1) (Appendix A, Figure A3). The 1995 and 2001 LP Pakistan isolates were almost identical to each other, and both shared 99% nucleotide sequence identity in a BLAST similarity search with the 1979 Northern Ireland sequence. It is therefore likely that the LP Pakistan isolates are contaminants, a conclusion which was also recently reached in another study of H7 HA (Lebarbenchon and Stallknecht 2011).

Only two HA sequences from the Netherlands H7N7 2003 outbreak (and one sequence from the spread of the outbreak to Germany) were available on the NCBI database at the time that this analysis was performed, although almost 200 HA, NA and PB2 sequences from this outbreak have recently been available on the GISAID website (http://platform.gisaid.org/epi3/frontend). The Netherlands H7N7 HA sequences fell within a clade of sequences obtained from wild waterfowl in Europe and Mongolia between 2000 and 2002, from viruses of subtypes H7N1, H7N3, H7N7 and H7N9. These sequences include an H7N3 isolate from a wild mallard in the Netherlands in 2002, and suggest that, like the Italian avian influenza outbreaks, the Netherlands H7N7 outbreak was caused by transmission of virus from wild to domestic birds, as had previously been suggested (Fouchier *et al.* 2004).

Two other, phylogenetically distinct, clades containing wild bird sequences are present in the Eurasian phylogeny. One clade consists of sequences sampled in East Asia between 2003 and 2009, from viruses of subtypes H7N2, H7N3, H7N7 and H7N9. The other clade consists of avian influenza sequences from wild birds isolated from European (Spanish, Portuguese, Hungarian, Slovenian and Ukranian) H7N1, H7N2, H7N3, H7N7, H7N8, H7N9 viruses between 2005 and 2009. The fact that these clades are contemporaneous suggests the presence of multiple co-circulating lineages of avian H7 HA in wild birds. A small number of sequences sampled from domestic birds in England, Denmark and the Czech Republic also appear in the clade of European wild bird sequences, again demonstrating the close genetic relationship between viruses in wild and domestic birds. Another notable clade which can be identified in the Eurasian BEAST MCC tree for avian H7 HA sequences contains LP H7N1 sequences isolated from caged and aviary birds (parrot, parakeet, conure, common iora and fairy bluebird) from Europe and Asia in 1994 and 1995. The size of the geographical region from which these genetically similar viruses were isolated (between England in the west and Hong Kong in the east) demonstrates the potential for the rapid spread of avian influenza through the transportation of domestic birds.

The North American BEAST MCC tree of avian influenza H7 HA sequences separates into two main lineages after 1990. One lineage consists of sequences isolated from domestic birds (from live bird markets and poultry farms) between 1994 and 2006. Except for two H7N3 sequences from 1994, all of the sequences in the clade were from viruses of subtype H7N2. The other major clade sampled after 1990 in the North American phylogeny consists mainly of sequence isolated from wild water fowl between 2001 and 2009, from viruses of subtypes H7N2, H7N3, H7N4, H7N6, H7N7 and H7N9. This clade also contained one HP sequence from a domestic chicken in Canada (A/Chicken/SK/HR00011/2007(H7N3)). Although there is evidence for geographical clustering by East and West coast within the clade of wild bird sequences, the separation is not complete and suggests that mixing does occur between these regions. Sequences from the British Columbia H7N3 outbreak of 2004-2005 can be observed in the North American avian influenza H7 HA phylogeny, falling basal to the clade of wild bird sequences.

Estimates of the rates of nucleotide substitution were obtained in BEAST for the North American dataset (4.44 x10^{-3} substitutions/site/year, 95% HPD = 3.79 x10^{-3}, 5.07 x10^{-3}) and the Eurasian/African dataset (4.54 x10^{-3} substitutions/site/year, 95% HPD = 3.51 x10^{-3}, 5.64 x10^{-3}). Including the pre-1990 sequences in the BEAST analysis of Eurasian sequences led to a slightly lower, albeit overlapping, substitution rate estimate of 3.90 x10^{-3} substitutions/site/year (95% HPD = 3.15x10^{-3}, 4.73x10^{-3}), possibly resulting from the position of the pre-1990 sequences far from the root of the tree. Overall, the substitution rate estimates are in line with the root-to-tip divergence plots, as well as previous estimates of the nucleotide substitution rate for avian influenza. For example, Chen and Holmes (2006) obtained mean substitution rate estimates of between 2x10^{-3} and 6x10^{-3} for 12 out of 13 avian influenza HA datasets using BEAST.

3.5.6 Clustering of H7 HA sequences with respect to host and virus properties

In addition to background viral NA subtype (Figure 3.7 and Figure 3.8), the tips of the BEAST MCC trees were coloured according to whether sequences were sampled from a wild or domestic host (Figure 3.9) and whether the virus was of low or high pathogenicity (LP or HP respectively) (Figure 3.10). In both the North American and Eurasian/African trees, it is possible to observe that avian H7 HA sequences do not cluster according to the background NA subtype of the virus. Instead, H7 sequences from viruses with different NA subtypes are distributed across the tree. Such a pattern is indicative of reassortment between influenza segments. This behaviour is particularly noticeable amongst the North American wild bird sequences from 2006, where the predominant serotype is H7N3, but sequences of serotypes H7N2, H7N4 and H7N7 are interspersed across the clade. Highly pathogenic viruses are distributed across the BEAST phylogenies and do not form a distinct phylogenetic lineage, consistent with previous reports (e.g. Rohm *et al*. (1995); Banks *et al*. (2000)). In addition, isolates from wild and domestic birds cluster together in various parts of the trees, suggesting repeated introduction of avian influenza viruses from wild birds into domestic poultry populations.

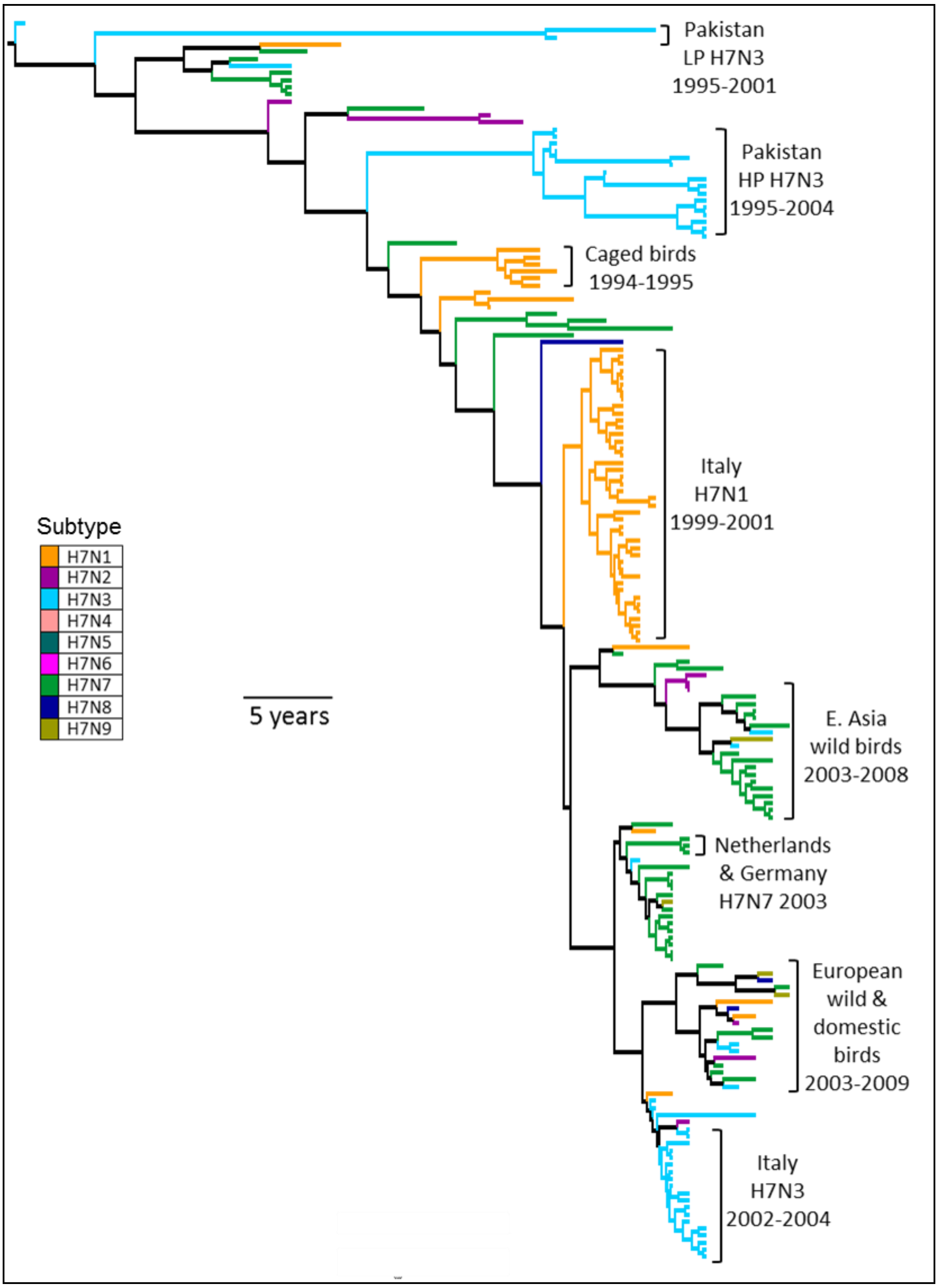


Figure 3.7
Time-scaled Bayesian phylogeny of Eurasian Avian H7 HA influenza sequences. Maximum clade credibility (MCC) tree constructed from posterior phylogeny samples obtained using the BEAST software. Analysis was performed under an SRD06 model of nucleotide substitution with a relaxed uncorrelated lognormal molecular clock. Branches at the tips of the phylogeny are coloured according to the background NA subtype of the virus for sequences at the tips of the tree (with basal branches in clades that are monophyletic with respect to subtype also coloured by subtype). Major outbreaks and clades are labelled on the phylogeny. Although the Pakistan LP isolates are included here, they were excluded from further BEAST analyses as discussed in Section 3.5.5.

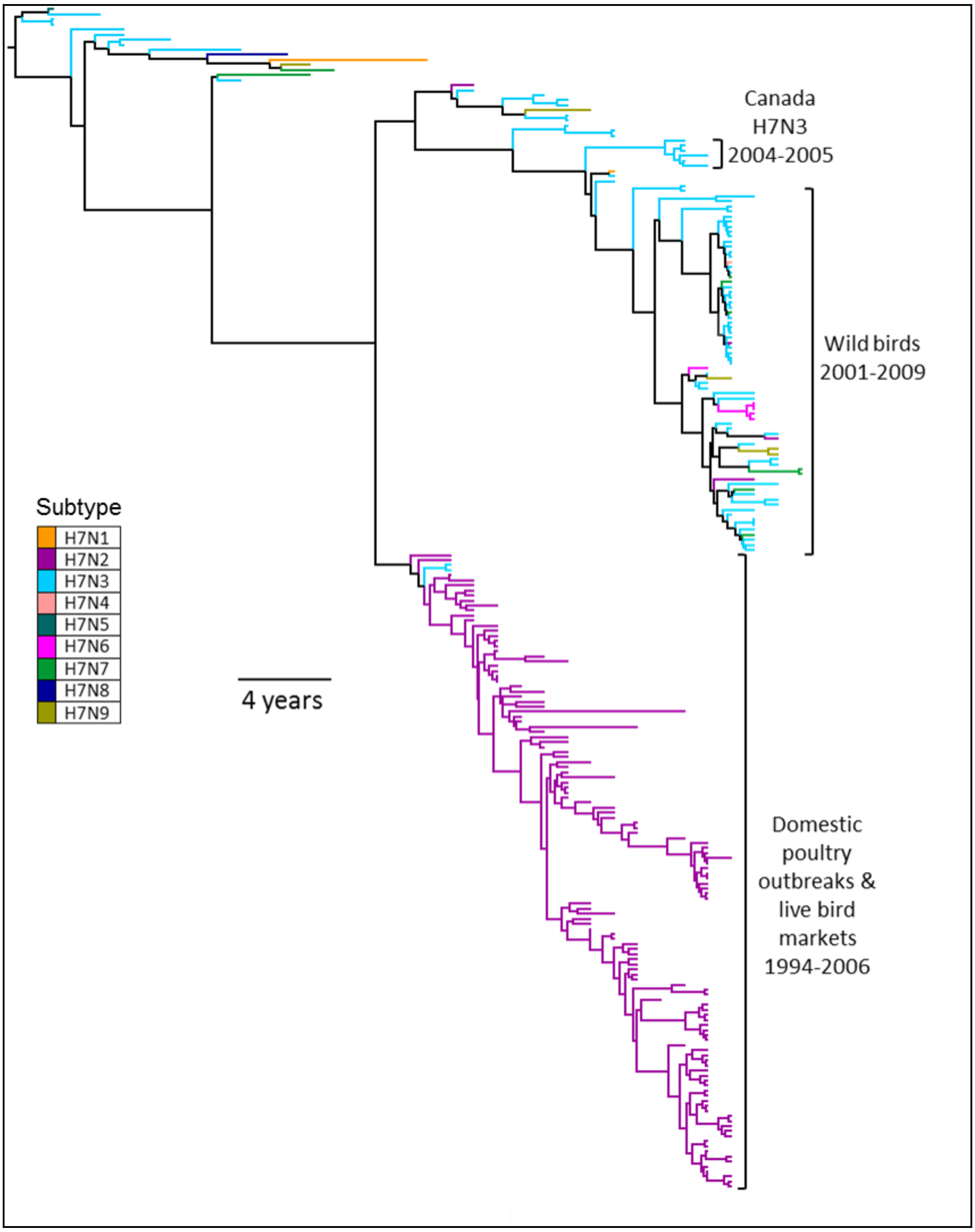


Figure 3.8
Time-scaled Bayesian phylogeny of North American Avian H7 HA influenza sequences. Maximum clade credibility (MCC) tree constructed from posterior phylogeny samples obtained using the BEAST software. Analysis was performed under an SRD06 model of nucleotide substitution with a relaxed uncorrelated lognormal molecular clock. Branches at the tips of the phylogeny are coloured according to the background NA subtype of the virus for sequences at the tips of the tree (with basal branches in clades that are monophyletic with respect to subtype also coloured by subtype). Major outbreaks and clades are labelled on the phylogeny.

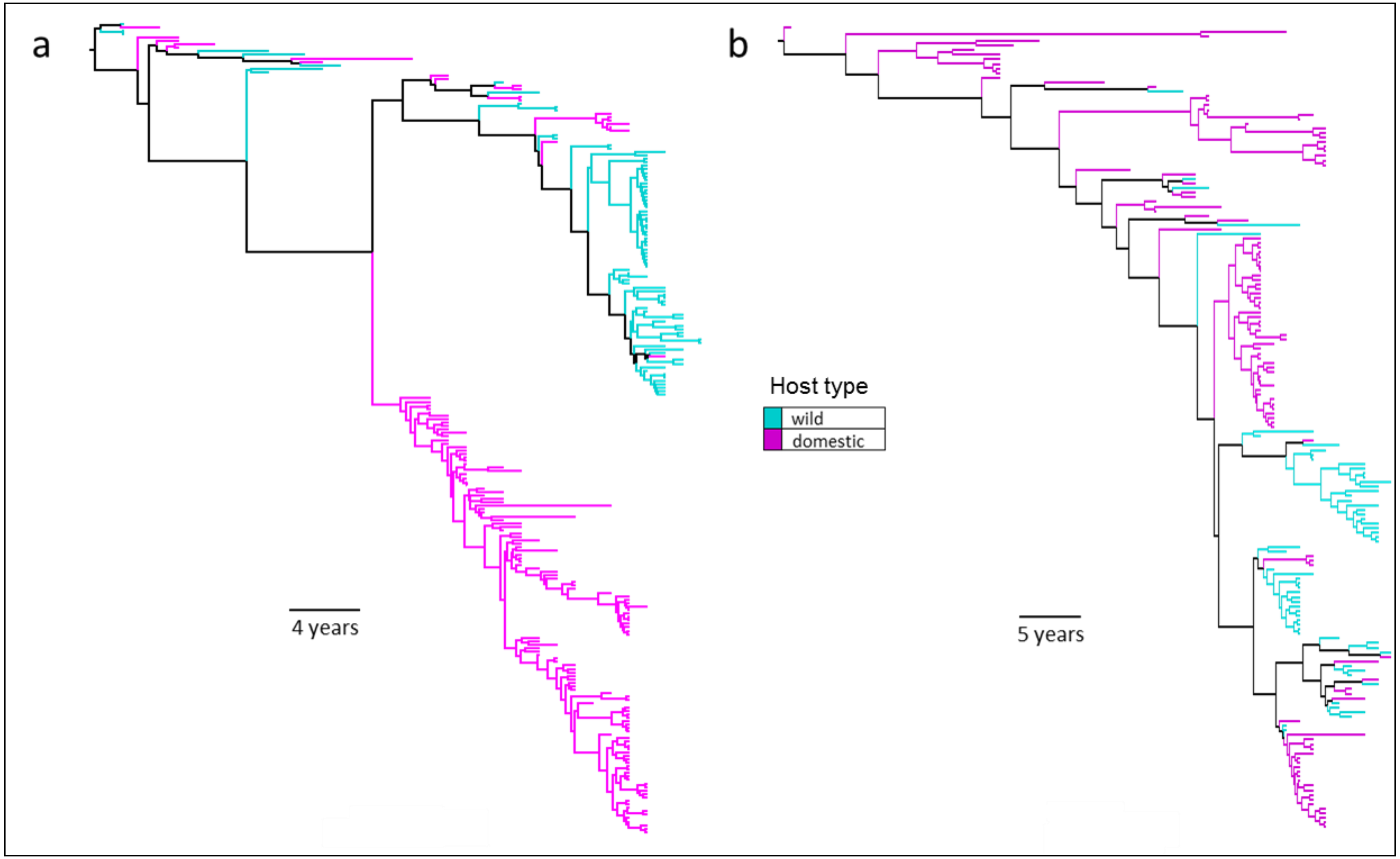


Figure 3.9: **BEAST avian H7 HA phylogenies coloured by host type.** Maximum clade credibility (MCC) trees for (a) North American and (b) Eurasian sequences, constructed from posterior phylogeny samples obtained using the BEAST software. Tips of the tree (and clades monophyletic with respect to host type) are coloured according to whether the host from which the virus was sampled was a wild or domestic bird.

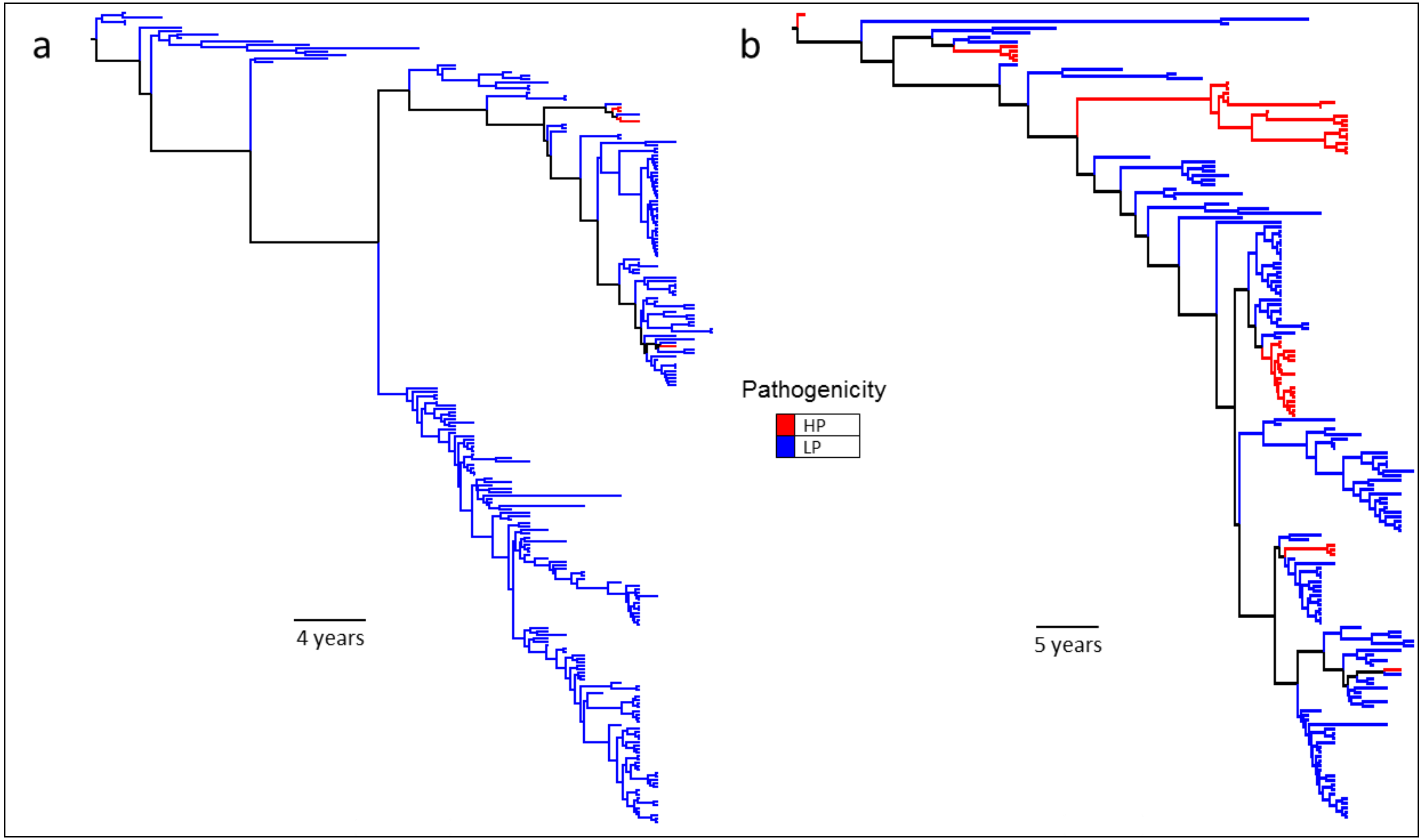


Figure 3.10: **BEAST avian H7 HA phylogenies coloured by pathogenicity.** Maximum clade credibility (MCC) trees for (a) North American and (b) Eurasian sequences, constructed from posterior phylogeny samples obtained using the BEAST software. Tips of the tree (and clades monophyletic with respect to pathogenicity) are coloured according to whether the virus was highly pathogenic (HP) or of low pathogenicity (LP).

3.6 Discussion

In this chapter, evolutionary analysis of full length H7 avian influenza HA sequences available from the NCBI Influenza Virus Resource was undertaken. The distribution of available sequences was reported, in terms of location and year of sampling, taxonomic order of avian host and NA subtype and pathogenicity of the virus. All observed cleavage site motifs were recorded, along with their frequency and the distribution of hosts and viruses in which they were found. Phylogenies were obtained which showed global evolutionary relationships between H7 avian influenza HA viruses, and analysis with BEAST allowed visualisation of trees on an explicit timescale. The distribution of different host or virus properties across the phylogenies was also considered.

Overall, a large number of avian influenza virus sequences were available, representing many combinations of host, location, pathogenicity and NA subtypes. However, it is not clear to what extent variation in frequencies (such as the low number of South American and African sequences compared to those in North America and Eurasia) is a reflection of the true distribution of avian influenza viruses or an artefact of differences in levels of surveillance and sampling. Substantial differences even exist between the numbers of sequences available for different outbreaks in domestic poultry in similar regions: for example, 40 HA sequences were available from the Italian H7N1 outbreak of 1999-2000, whereas only 3 HA sequences from the 2003 Netherlands H7N7 outbreak were available in the NBCI database. Strategies for subsampling sequences should perhaps be considered in future evolutionary analyses, to reduce bias resulting from variation in the intensity of sampling in different regions or avian populations.

Phylogenetic analyses using neighbour-joining and maximum likelihood methods, as well as Bayesian coalescent-based inference using BEAST, confirmed the existence of distinct North/South American and Eurasia/African/Australian lineages, with no evidence of transatlantic transmission of avian influenza viruses by wild birds in this dataset. This reinforces the hypothesis that transmission of avian influenza viruses is closely related to the migratory routes of wild birds (Olsen *et al.* 2006). The observation that three sequences sampled from caged birds in North America clustered with earlier or contemporaneous sequences from caged birds in the Eurasian/African clade indicates that transportation of caged birds could provide a mechanism for the global spread of influenza viruses. Trade, especially the smuggling of cage-birds, has previously been implicated as a major risk for the introduction of Eurasian avian influenza viruses into North America, rather than transmission via wild birds (Webster *et al.* 2007). For example, Webster *et al.* suggested that this is the most likely route through which HPAI H5N1 could enter North America from Eurasia.

The finding that the Australian avian H7 HA sequences form their own distinct lineage within the Eurasian/African/Australasian clade has recently been highlighted by Bulach *et al*. (2010). A strong positive correlation was observed between root-to-tip distance and date of sampling for Australian H7 HA sequences associated with numerous NA backgrounds (N2, N3, N4, N6 and N7) (Figure 3.6). This indicates the maintenance of an Australian avian influenza virus reservoir distinct from the Eurasian wild bird reservoir, with repeated reassortment within Australia, as suggested by Bulach *et al*. (2010). Similarly, a distinct South American clade was observed within the major North/South American lineage. The finding that HA sequences from the 2002 Chilean H7N3 outbreak in chickens clustered most closely with an H7N3 sequence from a wild Bolivian duck, A/cinnamon teal/Bolivia/ 4537/2001(H7N3), suggests the possibility of a distinct avian influenza virus reservoir in this under-studied region, with the potential to cause outbreaks in domestic poultry.

It is likely that the apparent genetic isolation of Australasian sequences within the Eurasian/African/Australasian lineage is a result of geographical isolation linked to the migratory flyways of wild birds (Olsen *et al.* 2006) (see Chapter 1, Figure 1.3). The divergence of the New Zealand sequence from those isolated in Australia and

Tasmania indicates the possibility of distinct separate avian influenza virus reservoirs within Australasia. Although distinct North and South American clades may be observed in the avian H7 HA phylogenies, latitudinal migration of wild birds does take place between North and South America (Olsen *et al.* 2006). However, without additional sequence information and analysis it is not possible to determine the extent to which mixing of H7 avian influenza viruses between North and South America takes place. Increased global surveillance of avian influenza in wild birds could help to shed light on these matters.

By surveying the available H7 avian influenza HA data, it was possible to provide a global picture of H7 avian influenza virus circulation over multiple decades and to identify subsets of sequences upon which further, more detailed, studies could be carried out. For example, many sequences are available from the 1999-2000 H7N1 Italian avian influenza outbreak in domestic poultry, where a low pathogenic virus mutated into a highly pathogenic form and was eradicated using culling and vaccination. In addition, the long-term evolution of avian influenza viruses in domestic hosts could be analysed using sequences from the H7N2 virus which circulated in the live bird markets of North America between 1994 and 2006. I also considered the distribution of different host or virus properties across the time-scaled avian influenza H7 HA phylogenies. This provided the basis for other studies in this thesis, including investigating the selective pressure experienced by H7 HA on different NA subtype backgrounds (Chapter 4) and using discrete trait mapping methods to quantify reassortment from phylogenies (Chapter 5).

Chapter 4

Evolutionary interactions between HA and NA in avian influenza

4 Evolutionary interactions between HA and NA in avian influenza

4.1 Chapter Summary

Reassortment between the RNA segments encoding haemagglutinin (HA) and neuraminidase (NA), the major antigenic influenza proteins, produces novel combinations of influenza subtypes and has preceded the emergence of pandemic influenza viruses. HA and NA have interacting roles in the viral life cycle, and are responsible for binding of virions to host cells and release of progeny virions from host cells respectively. It has been suggested that a balance between HA and NA activity is required for productive viral infection, and that genetic interactions between the segments encoding HA and NA could mediate this functional balance. In this chapter, I perform a Bayesian analysis to investigate how association with different NA subtypes influences the evolution of H7 HA, in terms of synonymous and non-synonymous substitution rates (d_S and d_N respectively) and their ratio (d_N/d_S). I use stochastic mutational mapping to infer codon changes along different parts of the H7 HA phylogeny, corresponding to lineages of different NA subtype backgrounds. This allows d_N, d_S and d_N/d_S to be calculated for H7 HA on each NA subtype background. The findings indicate that the selective pressure experienced by H7 HA may vary substantially depending on the associated NA subtype of the virus. Although it is difficult to exclude all possible confounding factors amongst the available data, no substantial difference in d_N/d_S was detected between viruses of high and low pathogenicity, or between lineages corresponding to different taxonomic orders of avian host.

4.2 Chapter Aims

- Investigate how selective pressure (measured by the ratio d_N/d_S) varies in H7 HA on different NA subtype backgrounds, and consider patterns of sites under putative positive selection
- Test for differences in selective pressure between HP and LP avian influenza viruses, as well as between viruses in ducks and chickens

4.3 Introduction

The main antigenic influenza proteins, the surface proteins haemagglutinin (HA) and neuraminidase (NA), are each encoded by a separate RNA segment and are classified into subtypes which do not cross-react serologically and are phylogenetically distinct. Whilst all HA and NA subtypes circulate in wild waterfowl (Webster *et al.* 2007), viruses with certain HA/NA combinations occur frequently in nature whereas others are rarely detected (Kaverin *et al.* 2000; Alexander 2003; Munster *et al.* 2007). This, combined with the failure of laboratory studies to produce viable reassortant viruses of particular subtype combinations, has led to the suggestion that there is a requirement for a functional match between the HA and NA subtypes (Wagner *et al.* 2002).

The HA and NA proteins play complementary roles in the life cycle of the influenza virus. Both HA and NA bind to host cell receptors containing sialic acid residues: HA to initiate viral entry into the host cell, and NA to permit the release of viral progeny from infected cells (see Chapter 1, Section 1.2). Experimental studies have suggested that a fine balance between HA and NA activity must be achieved for productive viral infection (Wagner *et al.* 2000). For example, if HA activity, and thus receptor-binding avidity, is high, then a high level of NA activity may also be required in order to prevent a reduction in virus yield resulting from the aggregation of progeny virions below the surface of host cells (Rudneva *et al.* 2003) (note that virus yield is measured by the number of progeny virions released by infected host cells). A balance between HA and NA activity may, in fact, be more important for viral fitness than high levels of activity *per se*. For example, Kaverin *et al*. (1998) showed that, when artificially generated reassortant viruses of the N1 NA subtype were cultured, several (e.g. H3N1) only gave low yields. However, when the low-yield H3N1 culture was passaged, a number of changes occurred in the HA which reduced its receptor binding activity, apparently to match that of the NA in the reassortant virus rather than to return to the high levels of HA activity found in the H3N8 parent virus. Adaptive post-reassortment changes in the receptor-binding region of the HA during serial passage have also been observed in other experimental studies using avian H2, H3, H4, H10 or H13 HA and low-activity human N1 NA (reviewed by Rudneva *et al*. (2003)).

Both the HA and NA proteins are thought to determine the sensitivity of naturally-occurring influenza viruses to neuraminidase-inhibiting drugs (NAIs) (Baigent *et al.* 1999). *In vitro* studies have investigated genetic interactions between HA and NA in terms of NAI resistance. Evidence suggests that mutations in the HA which decrease receptor binding avidity may compensate for a decrease in NA activity resulting from treatment with NAIs, thus restoring the balance between HA and NA function (Gubareva *et al.* 1996; McKimmBreschkin *et al.* 1996; McKimm-Breschkin *et al.* 1998; Wagner *et al.* 2002). In addition, HA and NA mutations which individually confer low-level resistance to NAIs have been found to combine synergistically to confer resistance at a higher level (Blick *et al.* 1998). Interdependence between the length of the NA stalk section and the number of HA glycosylation sites has been identified in laboratory strains (Wagner *et al.* 2000; Baigent and McCauley 2001) and may also have direct consequences for the transmission of influenza viruses to other host species. For example, influenza A viruses which have become established in terrestrial poultry may possess additional HA glycosylation sites, accompanied by deletions in the stalk section of their NA (Matrosovich *et al.* 1999; Banks *et al.* 2001).

Reassortment – the exchange of genetic segments between co-infecting parental viruses during replication – leads to novel combinations of influenza HA, NA and other segments. Reassortment has been implicated in the emergence of pandemic influenza viruses, including those of avian origin which were responsible for significant human mortality in the twentieth century (Scholtissek *et al.* 1978; Kawaoka *et al.* 1989) and the recent H1N1 pandemic strain (Smith *et al.* 2009). Naturally-occurring reassortment events could affect the functional balance between the HA and NA proteins (Wagner *et al.* 2002) and this could in turn affect their evolution. Whilst previous studies, such as those of Suzuki and Nei (2002) and Chen and Holmes (2006), have investigated evolutionary rates of influenza, few have focused on how rates of evolution are affected by genetic interactions between segments (Rambaut *et al.* 2008).

Evolution of protein coding sequences can be quantified in terms of rates of synonymous (d_S) and non-synonymous substitution (d_N) and their ratio, d_N/d_S, following the counting-based methods of Miyata and Yasunaga (1980), Li *et al.* (1985) and Nei and Gojobori (1986). The d_N/d_S ratio can be used to make inferences about the nature of the selective pressure acting upon a gene or a particular group of sites. Departures from selective neutrality can be detected when d_N/d_S differs from 1, with $d_N/d_S > 1$ indicating positive selection and $d_N/d_S < 1$ being indicative of purifying selection (see Chapter 2, Section 2.11). However, gene-wide estimates of d_N/d_S which show overall purifying selection may mask a small number of sites experiencing positive selection. Whilst the overall rate of non-synonymous substitution is lower than the synonymous substitution rate across the influenza HA in humans and swine (e.g. Sugita *et al*. (1991)), as well as in birds (e.g. Chen and Holmes (2006)), evidence has been provided for positive selection at certain sites, particularly those of antigenic significance (e.g. Fitch *et al*. (1991), Ina and Gojobori (1994), Bush *et al*. (1999), Kosakovsky Pond *et al*. (2008)).

In this chapter, a Bayesian stochastic mutational mapping approach (Nielsen 2001; Nielsen 2002) was used to investigate how the association with different NA subtypes influences the evolution of the HA-encoding segment of H7 avian influenza. The method of Nielsen (2001, 2002) was extended by rescaling counts of synonymous and non-synonymous changes to calculate d_N and d_S. Ancestral trait mapping was used to construct a model that assigned background NA subtypes to branches across the tree, and d_N/d_S was averaged across all parts of the tree corresponding to a particular NA subtype. The method accounts for a lack of monophyly across the tree with respect to NA subtype background, which arises through repeated exposure of H7 HA to different NA backgrounds via reassortment. This provides an advantage over studies such as that of Chen and Holmes (2006) who constructed separate datasets according to background NA subtype, despite the sequences being distributed across the avian H7 HA phylogeny. In this chapter, significant differences are observed between HA1-wide d_N/d_S for H7 avian influenza on different NA subtype backgrounds, consistent with the hypothesis that the selective pressure experienced by HA can be affected by the genetic context in which the segment finds itself.

4.4 Methods

4.4.1 Dataset

All available (as at April 2008) complete avian H7 nucleotide sequences for the HA coding region were downloaded from the NCBI database (www.ncbi.nlm.nih.gov) (Bao *et al.* 2008) and labelled according to the corresponding NA subtype of the virus. Sequences were screened for identity and, when identical nucleotide sequences were present, only one of the sequences was included. Only NA subtypes for which more than 20 sequences were available were analysed; these subtypes were N1 (62 sequences), N2 (75 sequences), N3 (69 sequences) and N7 (47 sequences). Standard names for the sequences used in this study are provided in Appendix B, Table B1.

Alignment of sequences was performed manually, using BioEdit (Hall 1999). The alignment of H7 HA sequences was split at the HA1/HA2 cleavage site (Perdue *et al.* 1997). The HA1 coding region, which is approximately two thirds of the length of the whole HA and has the major antigenic role for the virus (Palese and Shaw 2007), and the signal peptide region (17 amino acids immediately preceding the start of the HA1), were analysed (hereafter known as the HA1 alignment). After excluding sites at which there were gaps for a large proportion of sequences, the alignment was 987 nucleotides (329 codons) in length. A test for recombination using a single breakpoint analysis (Kosakovsky Pond *et al.* 2006a) in the HyPhy software (Kosakovsky Pond *et al.* 2005; Kosakovsky Pond *et al.* 2006b) found no evidence of recombination in the H7 HA alignment.

Where possible, the avian H7 HA sequences were also classified according to the taxonomic order of the avian host from which the virus was isolated (see Chapter 3, Table 3.1), and whether the virus was highly pathogenic (HP) or low pathogenic (LP), on the basis of literature searches and examination of the HA cleavage site motif (see Appendix A, Table A1).

4.4.2 MrBayes phylogenetic analysis

MrBayes version 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) was used to obtain posterior samples of topologies, branch lengths and substitution model parameters for the H7 HA1 alignment. An outgroup sequence (A/Australian shelduck/Western Australia/1756/1983(H15N2); GenBank accession no. ABB90704) was used to root the trees. H15 has previously been shown to be the closest subtype phylogenetically to H7 HA (Chen and Holmes 2006). A General Time Reversible (GTR) model of nucleotide substitution (Tavaré 1986) was selected, which allowed for unequal equilibrium base frequencies, different substitution rates between nucleotides and gamma-distributed rate variation across sites.

Three independent MrBayes runs were conducted, with Markov Chain Monte Carlo (MCMC) searching over 2,000,000 generations in each run. Trees and parameters were sampled every 1000 generations. The Tracer software (Rambaut and Drummond 2007) was used to inspect the chain traces, which indicated that a burn-in period of 1,000,000 generations was required before the chains had converged. The chain traces were compared across the three runs, and similar post-burnin values were observed in all runs. A post-burnin sample of 1000 posterior trees and sets of parameter estimates was obtained for further analysis.

4.4.3 Inferring mutational histories

Bayesian stochastic mutational mapping (Nielsen 2001; Nielsen 2002; Huelsenbeck *et al.* 2003) was used to infer mutational histories (maps), using the posterior phylogeny samples from MrBayes. Mutational histories describe the nature and locations of molecular changes along the branches of phylogenies (Figure 4.1). Stochastic mutational mapping is a Bayesian approach under which mutational histories are sampled from the posterior distribution of mutational mappings, given the observed nucleotide data (see Chapter 2, Section 2.10.3 for a detailed account of the mutational mapping process).

For each of the 1000 post-burnin MrBayes topology and substitution model samples, 10 mutational mappings were sampled for each nucleotide site in the H7 HA1 alignment; this accounted for phylogenetic uncertainty and the fact that there are multiple possible mutational paths along the tree. Within each phylogeny sample and mutational mapping replicate, the mutational history of each codon in the alignment was reconstructed by combining the mutational maps for the first, second and third codon positions. Branch lengths from the maps for codon positions 1 and 2 were rescaled to the branch length at codon position 3. This allowed codon substitutions to be identified (Figure 4.2). The number of synonymous and non-synonymous changes (C_s and C_n respectively) along different parts of the tree, and the timings of the changes, were recorded. If a mutational history was sampled such that a stop codon would occur at some point along the mutational path, the nucleotide maps associated with this codon map were rejected and re-sampled until a map with no stop codons was obtained.

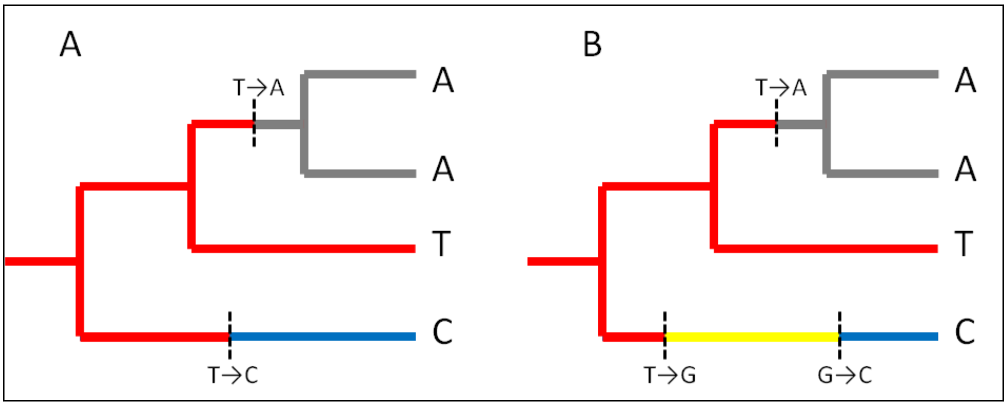


Figure 4.1
Example nucleotide mutational maps. The stochastic mutational mapping process is used to infer mutational histories for nucleotide sites, which report the nature and location of molecular changes along a phylogeny. Multiple mutational mappings may be sampled for each site. For example, maps (A) and (B) are both valid reconstructions for the observed pattern of variation. Note that map (B), where there are three nucleotide changes, would not be permitted under the parsimony criterion, since the minimum number of changes required to fit the nucleotide data is two, as in map (A).

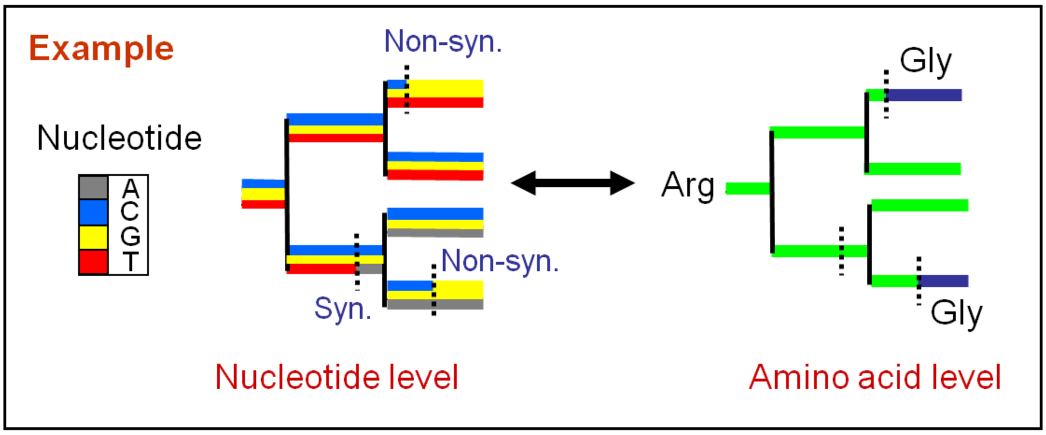


Figure 4.2
Example codon map obtained using stochastic mutational mapping. For each codon site, the first and second codon position nucleotide maps for a site were rescaled to the branch lengths of the third position map and combined to produce a map at the amino-acid level. Nucleotide changes could then be labelled as synonymous or non-synonymous for calculating d_N, d_S and d_N/d_S. In this example there are three nucleotide changes, one of which is synonymous (CGT → CGA) and two of which are non-synonymous (CGT → GGT and CGA → GGA).

4.4.4 Scaling counts of non-synonymous and synonymous changes to calculate d_N and d_S

The method employed in this chapter extends the stochastic mutational mapping approach of Nielsen (2001, 2002) by rescaling observed numbers of synonymous and non-synonymous changes to account for differences in the evolutionary potential for synonymous or non-synonymous change at each codon position (i.e., the number of synonymous and non-synonymous sites in a specific codon). This is necessary because of the degeneracy of the genetic code. For example, there are two possible synonymous single nucleotide changes from the leucine codon TTA (TTG and CTA also code for leucine), whereas there are four synonymous sites for another leucine codon, CTA (CTT, CTC, CTG and TTA). The method also weights by the 'dwell time' – the time along the branch spent in each codon – to account for the fact that a higher number of changes would be expected over a longer period of evolutionary time than over a shorter period. The rescalings detailed below provide an expected value of d_N/d_S =1 under selective neutrality. Note that the assumption that synonymous changes are neutral, i.e. that there is no selective constraint from the RNA secondary structure, and no codon usage bias, is also being made here (see Chapter 2, Section 2.11.1), although both have been suggested to affect influenza virus evolution (Wong *et al.* 2010; Moss *et al.* 2011). For each codon site in the alignment, estimates of the number of synonymous and non-synonymous sites were calculated for a given part of the tree as follows:

$$S_s = \frac{1}{V_T} \sum_{i=1}^{c} \sum_{j=1}^{3} s_{ij} v_{ij}$$

$$S_n = \frac{1}{V_T} \sum_{i=1}^{c} \sum_{j=1}^{3} n_{ij} v_{ij}$$

where c = number of codon intervals (distinct codon states) along a part of the tree. A new interval occurs every time there is a nucleotide change, even if it is silent, since this alters the codon state

j = position of nucleotide site in the codon (1, 2 or 3)

s_{ij} = proportion of possible changes at the j^{th} codon position of the codon at interval *i* which are synonymous

n_{ij} = proportion of possible changes at the j^{th} codon position of the codon at interval *i* which are non-synonymous

v_{ij} = 'mutational time interval' or 'dwell time'. This is obtained by multiplying the substitution rate r_j with the length along the branch spent in each codon state. The parameter r_j is drawn from a gamma distribution, whose parameters were sampled during the MrBayes analysis. A value of r_j is sampled for each codon position (j = 1, 2, or 3) from its respective posterior distribution and the stochastic mutational map is then sampled under this rate

V_T = sum across all codon positions and over all codon intervals of the v_{ij}s,

$$V_T = \sum_{i=1}^{c}\sum_{j=1}^{3} v_{ij}.$$

Together with the v_{ij}s, this gives a time-weighted average which assigns more weight to codons with longer dwell times.

Note that, for a single codon interval, if the dwell time information is not used (i.e., if it is assumed that v_{ij} =1 for all *i* and *j*) then the calculation of the number of synonymous and non-synonymous sites is the same as that of Nei and Gojobori (1986), since s_{ij} here is equivalent to their f_i. However, unlike the Nei and Gojobori approach, by using the dwell time weighting the method here accommodates variation in branch lengths. Note also that Nei and Gojobori used the evolutionary distance formula of Jukes and Cantor (1969) to estimate the expected number of synonymous changes per synonymous site (or non-synonymous changes per non-synonymous site) from the proportions of synonymous and non-synonymous differences between pairs of sequences. However, the method employed in this chapter samples the full nucleotide state history across the phylogeny for each nucleotide site in the alignment, thus d_N and d_S may be estimated directly by counting synonymous and non-synonymous changes along branches and rescaling by numbers of synonymous and non-synonymous sites, and dwell times, as described above. The rescalings described above have now been implemented in SIMMAP version 1.5 (http://www.simmap.com/), released online in February 2010.

Values of C_s, C_n, S_s and S_n were used in calculating synonymous and non-synonymous evolutionary rates (d_S and d_N respectively) along different parts of the phylogeny samples, corresponding to background NA subtypes N1, N2, N3 and N7. In order to calculate d_N and d_S for H7 HA1 on different NA subtype backgrounds, parsimony mapping was used to assign ancestral NA subtypes at internal nodes along the MrBayes phylogeny samples, based on assignments at the tips of the phylogeny (i.e., the NA subtypes corresponding to the H7 HA sequences in the dataset) (Figure 4.3). This allowed branches to be classified by NA subtype: N1, N2, N3 or N7. Branches where a subtype could not be unambiguously assigned from a single pass of the parsimony algorithm from the tips of the tree to the root were not used in the analysis. This avoids the possible confounding factor of incorrect lineage classification which could arise from using methods which force ancestral states to be inferred for every branch, although the exclusion of ambiguous lineages potentially results in a loss of information. S_s and S_n were calculated across all branches to which a particular NA subtype had been assigned, and numbers of synonymous and non-synonymous changes were counted along those parts of the tree.

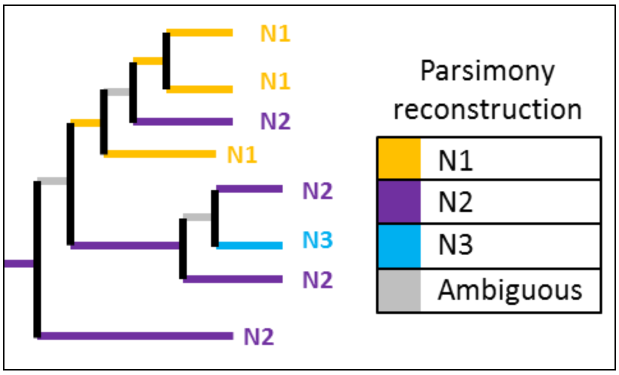


Figure 4.3
Example parsimony reconstruction of background NA subtypes on a phylogeny of H7 HA sequences. Branches are coloured according to the inferred ancestral subtype of the node immediately preceding them towards the tips of the tree. In the analysis performed in this chapter, a single-pass algorithm was implemented, which labels some branches as 'ambiguous'. This avoids the problem associated with erroneous assignment of subtypes in the subsequent calculation of evolutionary rates along branches associated with a particular NA subtype.

The rate of synonymous (d_S) change and the rate of non-synonymous (d_N) change were calculated as:

$$d_S = \frac{1}{T} . \frac{C_s}{S_s}$$

and

$$d_N = \frac{1}{T} . \frac{C_n}{S_n} .$$

.

Here, *T* was obtained by summing the branch lengths at all nucleotide positions in the codon, with branch lengths for the first and second codon positions rescaled to the third codon position lengths (i.e. 3* sum of the third position branch lengths), for all branches in the phylogeny to which a particular NA subtype had been assigned. Rescaling by the length of the portion of the tree corresponding to each background NA subtype allowed for a comparison of evolutionary rates between clades of different sizes. This extended previous mutational mapping approaches of Nielsen and others (Nielsen 2001; Nielsen 2002; Huelsenbeck *et al.* 2003), including those implemented in the SIMMAP software (Bollback 2006). By performing these calculations upon each of the 1000 MrBayes posterior phylogeny samples, approximations to the posterior distributions for d_N and d_S were obtained for each background NA subtype, at each codon site in the H7 HA1 alignment.

4.4.5 Post-processing of mutational mapping output

Python scripts were written to extract information from the raw mutational mapping output files and to average over the 10 mutational mapping replicates for each phylogeny sample. Estimates of d_N and d_S, obtained at each codon site for each background NA subtype, were averaged over the 10 mutational mapping replicates. Average values of d_N across the sites in the HA1 alignment were obtained for each background NA subtype by calculating the mean of the d_N values across all codon sites in the alignment (and similarly for d_S). For all 1000 MrBayes phylogeny samples, the average d_N estimate across all sites for a given NA subtype was divided by the corresponding d_S value for that subtype across all sites, to obtain an approximation to the posterior distribution for the H7 HA1 d_N/d_S for that subtype.

Estimates of d_N/d_S at individual codon sites in the H7 HA1 alignment were also calculated for each NA background subtype. For each site, d_N and d_S values were averaged over the 10 mutational mapping replicates for each tree, and then averaged over the 1000 MrBayes tree samples. To calculate the d_N/d_S ratio on a site-by-site basis, d_N for each site was divided by the average d_S value across the 329 codons for that subtype. The gene-wide d_S was used to avoid inflation of d_N/d_S values as a result of unobserved synonymous change at individual sites. Sites with an observed value of d_N/(gene-wide d_S) > 1 were identified as being under putative positive selection. Sites were converted from H7 to H3 numbering prior to being reported, as is the convention for influenza, and numbering was based upon the alignment of Nobusawa *et al*. (1991) (sites numbered -17 to -1 for the signal peptide region and 1 to 329 for the HA1 coding region). The HA1 domain in which putatively positively selected sites were found was reported using the alignment of Yang *et al*. (2010), in which portions of the HA corresponding to the fusion domain, vestigial esterase domain and receptor binding domain were identified.

4.4.6 Comparing posterior distributions of rates

Posterior distributions of d_N/d_S and rates of synonymous and non-synonymous substitution of avian H7 HA1 could be compared across background NA subtypes by considering highest posterior density (HPD) intervals (see Chapter 2, Section 2.1.2). A custom R script was written for plotting HPD intervals in a format analogous to confidence interval boxplots. After checking the distributions for unimodality, HPD intervals were calculated using the Chen and Shao algorithm (Chen and Shao 1999) in the *boa* R package for the analysis of Bayesian output (Smith 2007). The size of the overlap of the HPD intervals can be used as an indicator of whether the means of the distributions are significantly different.

In order to assess size of the overlap between posterior distributions of evolutionary rates of H7 HA on different background NA subtypes, the following comparison was implemented using ‘distributions of differences’. For rate distributions corresponding to arbitrary background NA subtypes A and B, multiple pairings of evolutionary rate estimates were drawn randomly from across the 1000 posterior samples, with one observation from subtype A and one from subtype B in each pair. The proportion of pairings for which the observed rate from subtype A was greater than the observed rate from B (and *vice versa*) was recorded. For a null hypothesis that there is no difference between the distributions, the point of interest is where zero lies in the distribution of paired differences. If the distributions for A and B were identical then the corresponding distribution of paired differences should be centred on zero, as one would expect A>B for half of the paired samples and A<B for the other half. However, if the proportion of samples for which A>B is extremely skewed (e.g. less than 0.05 or greater than 0.95) then zero lies in the tail of the distribution of paired differences, providing evidence that the location of the distributions is different (Figure 4.4). A total of 10^6 random pairings were sampled for each comparison of evolutionary rate distributions between different NA backgrounds. This yielded results which were identical to 2 significant figures to those obtained by systematically comparing each of the 1000 observations for one subtype with each of the 1000 observations for the other subtype. The values reported in this chapter are from the randomized pairing approach.

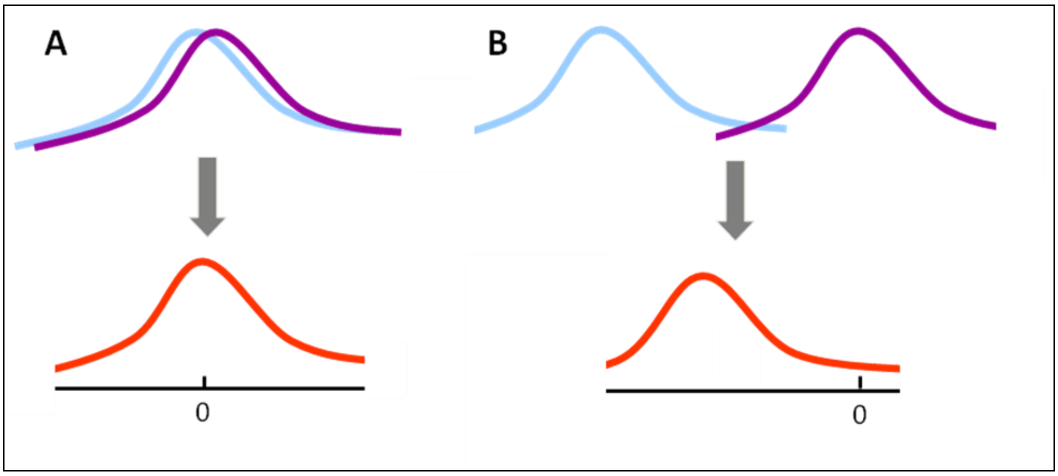


Figure 4.4
Testing for differences between posterior distributions of evolutionary rates for different NA background subtypes. (A) When the locations of the distributions (examples shown here in blue and purple) are very similar, the distribution of differences of randomised pairings between them (shown in red) will be roughly centred on zero. **(B)** When the distributions differ in their location, the distribution of differences between randomised pairings will be skewed, with zero at one of the tail ends. The proportion of pairings lying to each side of zero thus provides a measure of the difference in location of the distributions.

4.4.7 Assessing the effect of host type and pathogenicity

In this study, avian H7 HA sequences were labelled according to the NA subtype of the virus and rates of evolution were calculated for lineages corresponding to different NA backgrounds. In order to test whether a non-uniform distribution of host species or pathogenic viruses across different NA backgrounds could be confounding the ability to infer differences in d_N/d_S between subtypes, two further analyses were performed in an analogous manner to the NA subtype analysis. These analyses involved labelling sequences and performing stochastic mutational mapping to calculate and compare d_N/d_S between (a) HP and LP viruses and (b) viruses from different avian host orders. Bird orders compared were Galliformes (turkeys, chickens etc.) and Anseriformes (ducks, geese, etc.), with all other avian host orders combined (classified as ‘Other’) due to a paucity of sequences from these orders.

4.5 Results

4.5.1 Descriptive analysis of dataset

The dataset (Table 4.1 and Appendix B, Table B1) analysed in this chapter comprised 253 avian influenza H7 HA sequences from viruses with four different NA background subtypes: N1 (62 sequences), N2 (75 sequences), N3 (69 sequences) and N7 (47 sequences). Sequences were also classified according to the pathogenicity of the virus and the taxonomic order of the avian host from which the virus was isolated. Note this dataset of avian H7 HA sequences differed slightly from that described in Chapter 3 because it was an older download and rarer NA subtype backgrounds were excluded. Overall, 71% of the sequences were known to have been isolated from terrestrial poultry and approximately 16% were from aquatic fowl. Most of the sequences from Anseriformes were likely to have been isolated from farmed ducks (e.g. isolates labelled ‘duck’, with no additional information) although a small number were known to be from wild aquatic birds. On all NA backgrounds, the majority (over 63%) of sequences were from Galliformes, although isolates from Anseriformes were present for all subtypes (6 sequences from Anseriformes for H7N1 and H7N2; 13 for H7N3 and H7N7). Approximately two-thirds of the sequences were from HP viruses, although numbers of HP and LP isolates were not distributed evenly across the subtypes. Notably, H7N2 viruses have thus far only appeared in a low pathogenic form, whilst approximately half of the H7N1 isolates were from HP viruses.

For each background NA subtype, the H7 HA sequences covered a time-span of at least 25 years. There were roughly equal numbers of sequences from Eurasia and America (132 and 107 respectively), and isolates from Europe, Asia and North America were present for all four subtypes considered. The geographic spread of viruses of different subtypes appeared to differ between continents. For example, 85% of the H7N1 sequences and 74 % of the H7N7 sequences were from Europe, whilst 88% of the H7N2 isolates were from North America. As was observed in Chapter 3, H7N3 appeared to be the most ubiquitously sampled subtype, in terms of geographical location, avian host order and pathogenicity. Overall, geographic and temporal diversity appeared to be captured in all subtypes, although this information is not explicitly used in the phylogenetic analysis described in this chapter.

	Subtype			
All subtypes (253)	**H7N1 (62)**	**H7N2 (75)**	**H7N3 (69)**	**H7N7 (47)**
Host order				
Ans. (38)	Ans. (6)	Ans. (6)	Ans. (13)	Ans. (13)
Gal. (173)	Gal. (39)	Gal. (60)	Gal. (52)	Gal. (22)
Pathogenicity				
HP (56)	HP (20)	HP (0)	HP (20)	HP (16)
LP (195)	LP (42)	LP (75)	LP (49)	LP (29)
Time-span	1934-2001	1978-2006	1963-2006	1927-2003
Location				
Europe (118)	Europe (53)	Europe (5)	Europe (25)	Europe (35)
Asia (14)	Asia (4)	Asia (4)	Asia (3)	Asia (3)
Africa (4)	Africa (3)	Africa (0)	Africa (0)	Africa (1)
Australia (10)	Australia (0)	Australia (0)	Australia (4)	Australia (6)
N. America (99)	N. America (2)	N. America (66)	N. America (29)	N. America (2)
S. America (8)	S. America (0)	S. America (0)	S. America (8)	S. America (0)

Table 4.1
Composition of avian H7 HA sequence dataset (background NA subtypes N1, N2, N3 and N7 only). Numbers of sequences from different avian host taxonomic orders (Anseriformes = Ans., Galliformes = Gal.) and viral pathogenicities are provided, in addition to the time-span over which the sequences were sampled and the location of sampling. Numbers of sequences are given in brackets. Note that it was not possible to determine such information for all sequences.

4.5.2 Phylogenetic analysis

Consensus phylogenies obtained with MrBayes revealed similar patterns of clustering to the neighbour-joining, maximum likelihood and Bayesian (BEAST) analyses for avian H7 HA which were described in Chapter 3. In particular, it was possible to observe that H7 HA sequences clustered into major geographic lineages (Eurasian, American etc.) rather than forming distinct lineages according to the NA subtype, host order or pathogenic status of the virus (Appendix B, Figure B1).

Within the Eurasian clade, the Australian sequences and early European fowl plague viruses (sampled in the 1920s-1940s and abbreviated as FPV) both formed sub-lineages (each with posterior probability = 1) with long ancestral branches and appeared to be divergent from the later European, African and Asian sequences, as was observed in the ML and NJ analysis of Chapter 3. The NJ analysis of Chapter 3 also supported the phylogenetic position of FPV as a sister lineage to the Australian sequences, as has been observed in previous studies (Banks *et al.* 2000; Bulach *et al.* 2010; Lebarbenchon and Stallknecht 2011). The relatively low posterior probability (0.54) observed for the Eurasian clade could result from difficulty in placing the highly divergent FPV clade. As in Chapter 3, the FPV sequences were excluded from the analysis of evolutionary rates, since they have been extensively cultured and may show artificially high rates of molecular change.

On a smaller geographic scale, H7 HA sequences from within the same avian influenza outbreak clustered together in the MrBayes consensus trees. In line with Chapter 3, avian H7 HA sequences did not form distinct clades according to the NA background of the virus in the MrBayes consensus trees, which is indicative of repeated reassortment between HA and NA. Avian influenza H7 HA sequences also did not form distinct clades corresponding to HP or LP viruses, or the avian host orders from which they were sampled (Anseriformes, Galliformes or others).

4.5.3 Comparing evolutionary rates for H7 avian influenza HA across different NA background subtypes

The stochastic mutational mapping analysis yielded 1000 post-burnin posterior estimates of d_N and d_S for H7 avian influenza HA1 on each of background subtypes N1, N2, N3 and N7. The collection of estimates provides an approximation to the posterior distribution of d_N and *ds* for each subtype, and HPD plots for d_N and *ds* allowed posterior distributions of evolutionary rates to be calculated and visualised for H7 viruses with different NA subtypes (Figure 4.5 and Table 4.2). For each background NA subtype, the d_N value for each tree sample was divided by the d_S value for that tree sample to obtain approximations to the posterior distribution of

d_N/d_S. HPD plots of d_N/d_S for H7 HA1 on N1, N2, N3 and N7 NA backgrounds are also shown in Figure 4.5. The distributions of synonymous substitution rates (d_S) were very similar for H7 HA1 across all NA backgrounds, whilst the rate of non-synonymous substitution appeared to be higher for H7N2 than for H7N1, H7N3 or H7N7. The elevated rate of non-synonymous substitution for H7N2, in the absence of differences in synonymous substitution rates between the subtypes, led to an increased d_N/d_S for H7N2 compared to H7N1, H7N3 and H7N7. Although the mean of the H7N2 d_N/d_S samples lay outside of the 90% HPD intervals for the other subtypes, the lower limit of the H7N2 HPD interval overlapped slightly with the upper HPD limits of the other subtypes. It may also be noted that the rate of synonymous substitution was substantially higher than the rate of non-synonymous substitution for all subtypes, with no overlap between any of the HPD intervals for d_N and d_S. This resulted in gene-wide d_N/d_S estimates which were substantially less than one, indicating an overall pattern of purifying selection across the HA1.

In order to quantify differences in evolutionary rates for H7 HA1 on different NA backgrounds, comparisons of paired samples from different background NA subtypes were performed as described in Section 4.4.6. For subtypes A and B, the proportion (denoted **p**) of the paired samples for which the rate of subtype A was greater than for subtype B (the top value in each cell) or less than for subtype B (the bottom value in each cell) was reported. A split at least as extreme as 0.05/0.95 in either direction was interpreted as a substantial difference in the location of the distributions for the two subtypes. For all subtype comparisons, the distributions of paired differences for rates of synonymous substitution were roughly centred on zero (proportions approximately 0.5/0.5 greater than and less than zero), indicating no significant differences between the distributions, as suggested by the HPD plot. However, the pairwise difference comparisons indicated an elevated rate of non-synonymous change in H7N2, leading to a substantially higher d_N/d_S for H7N2 than for the other subtypes (**p** = 0.021 against H7N1; **p** = 0.009 against H7N3; **p** = 0.038 against H7N7) (Table 4.3).

Within MCMC tree samples, the relationship between d_N and d_S estimates was investigated for each subtype. In Figure 4.6, the d_N value for a tree sample was plotted against the d_S value for that tree sample, for each of the 1000 posterior rate estimates for each subtype. Within tree samples, a positive correlation between d_N and d_S was observed for all four subtypes using Pearson's product-moment correlation coefficient (H7N1: $r = 0.479$, $p < 0.001$; H7N2: $r = 0.579$, $p < 0.001$; H7N3, $r = 0.599$, $p < 0.001$; H7N7: $r = 0.562$, $p < 0.001$, with N=1000 points for each subtype[2]), indicating that a phylogeny sample with a higher rate of synonymous substitution would also have a higher rate of non-synonymous substitution, although the rate of synonymous substitution was not an exact predictor of the corresponding non-synonymous substitution rate. Positive correlation between gene-wide d_N and d_S has previously been observed amongst MrBayes phylogeny samples for HIV (Andrew Leigh Brown and Gareth Hughes, personal communication). In this study, such a correlation should not be due to a failure to correct for branch lengths, since these were explicitly incorporated into the rescaling of d_N and d_S values.

[2] The p-value from the Pearson product moment correlation coefficient corresponds to the null hypothesis that the correlation coefficient is zero. Assuming that the data are normally distributed, the sampling distribution for the correlation coefficient approximately follows a t-distribution with N-2 degrees of freedom (with N = number of pairs in the sample). Observed and theoretical quantiles were compared using normal Q-Q plots and indicated that the data were approximately normally distributed (not shown).

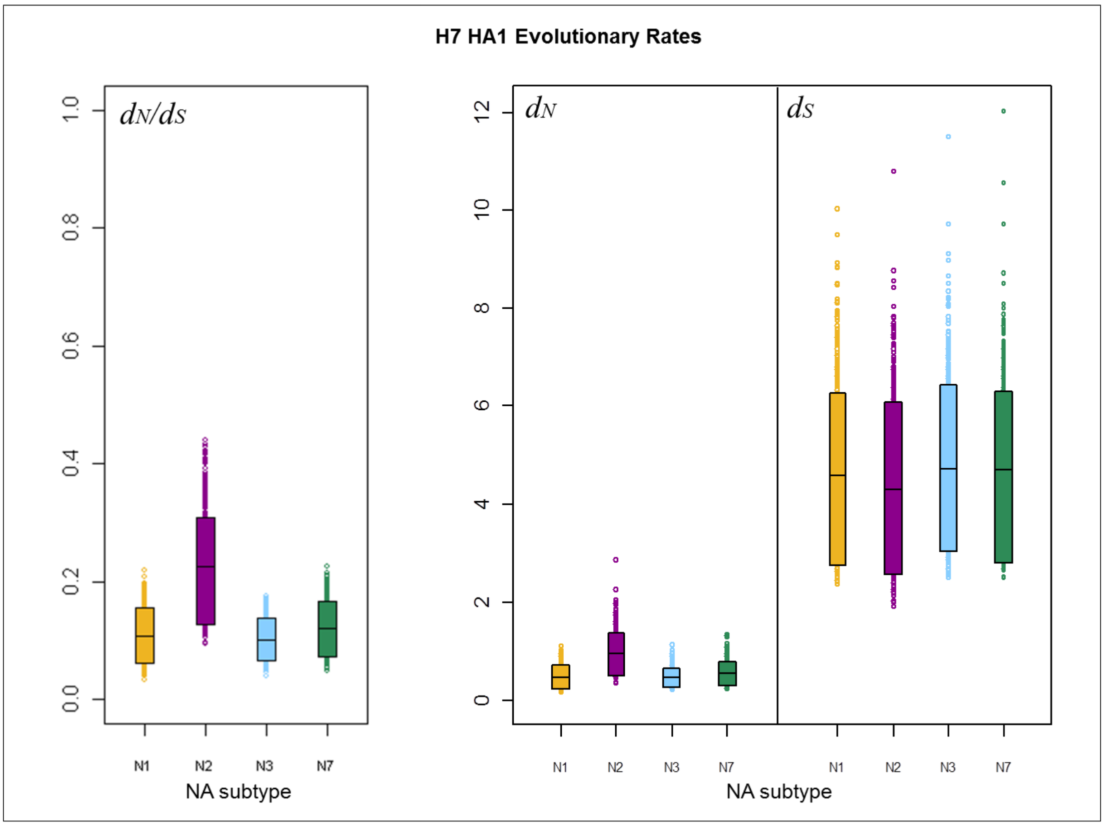


Figure 4.5

90% HPD plots for d_N/d_S, d_N and d_S for H7 HA1, split by NA subtype. The boxes show the limits of the narrowest interval containing 90% of the estimates. The horizontal lines inside the boxes indicate the location of the mean for each subtype. Individual points shown outside the boxes are values which lie below the lower limit, or above the upper limit, of the HPD interval. For each subtype, values for d_S are the number of synonymous changes per synonymous site, scaled by the total branch lengths in the tree sample for lineages corresponding to that subtype. Similarly, d_N is given in terms of the number of non-synonymous changes per non-synonymous site, scaled by the total branch lengths in the tree sample for lineages corresponding to that subtype.

Subtype	Mean	Lower 90%	Upper 90%
	d_N/d_S	HPD limit	HPD limit
H7N1	0.107	0.063	0.156
H7N2	0.226	0.126	0.309
H7N3	0.102	0.067	0.137
H7N7	0.120	0.074	0.168

Table 4.2
Average d_N/d_S across the H7 avian influenza HA1 on different NA backgrounds. For each background NA subtype, the average d_N/d_S across the HA1 coding region was obtained for each MCMC sample by first averaging over mutational mapping replicates on that tree, then calculating average values for d_N and d_S across all HA1 sites. Within tree samples, the site-averaged d_N was divided by the site-averaged d_S for that NA subtype, to obtain 1000 posterior estimates of the d_N/d_S ratio for each NA subtype background.

Comparison	d_N/d_S	d_N	d_S
H7N1-H7N2	0.021465 0.978535	0.048604 0.951396	0.577697 0.422303
H7N1-H7N3	0.540547 0.459453	0.503311 0.496689	0.467995 0.532005
H7N1-H7N7	0.373000 0.627000	0.356954 0.643046	0.468392 0.531608
H7N2-H7N3	0.991065 0.008935	0.965327 0.034673	0.389154 0.610846
H7N2-H7N7	0.962234 0.037766	0.907221 0.092779	0.390056 0.610846
H7N3-H7N7	0.317627 0.682733	0.340218 0.659782	0.501494 0.498506

Table 4.3
Comparing posterior distributions of evolutionary rates for avian influenza HA1 across different background NA subtypes using randomised pairings. The proportion of randomised pairings of posterior rate samples for which the value for the first subtype in the comparison, minus the value for the second subtype in the comparison, was greater than 0 (top value in each cell) and less than 0 (bottom value in each cell) was reported. Similar distributions would be indicated by the difference being greater than 0 (likewise less than 0) in approximately 50% of pairings. Differences in the location of the distributions would be indicated by a more extreme split in one direction (cells highlighted in yellow indicate a split at least as extreme as 5%, and cells highlighted in orange indicate a split of between 5% and 10%). All comparisons involving H7N2 indicated that this subtype had a higher d_N/d_S ratio than the other subtypes.

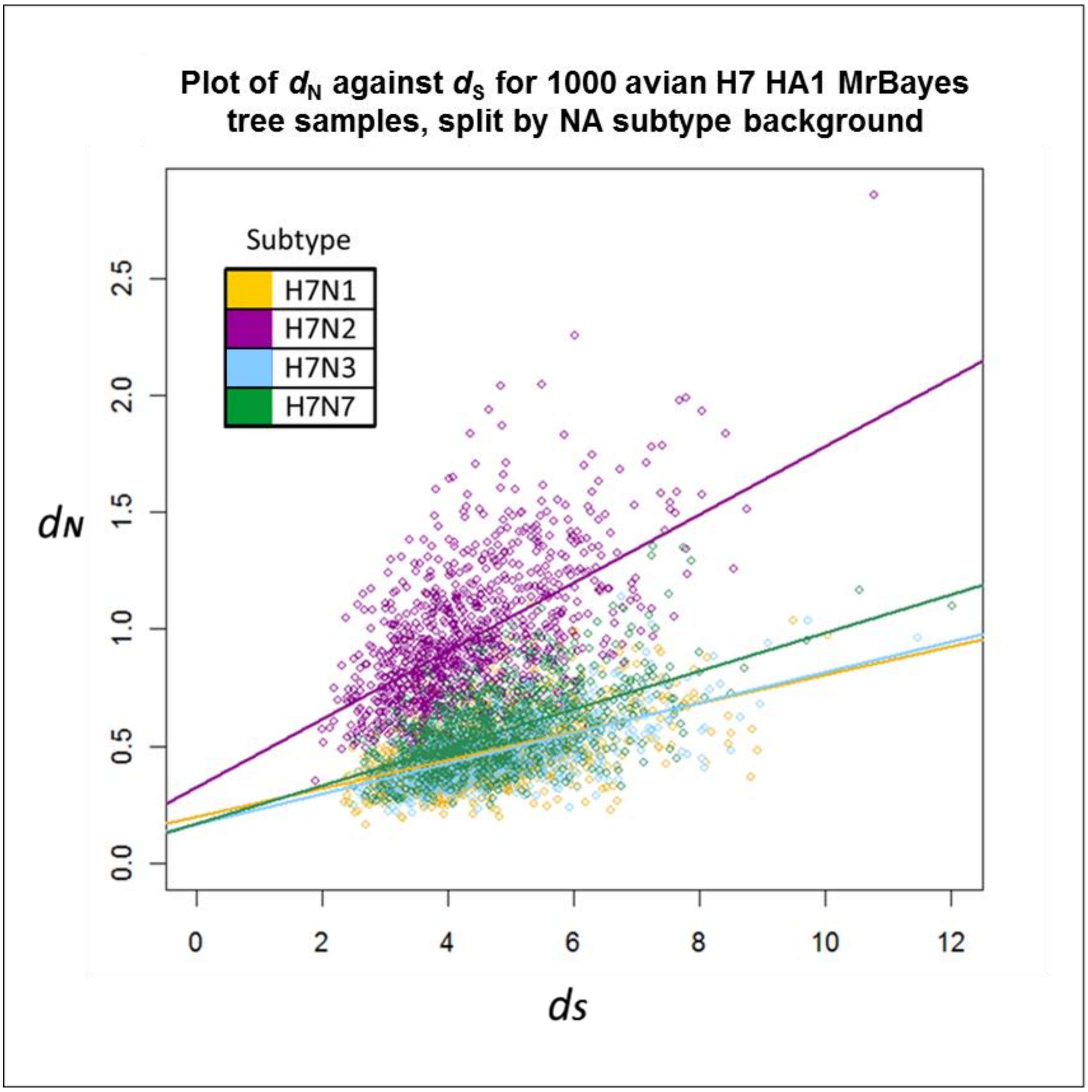


Figure 4.6

The rate of non-synonymous substitution (d_N) plotted against the rate of synonymous substitution (d_S) for avian influenza H7 HA1 from viruses with different background NA subtypes. For each of the 1000 MCMC tree samples from MrBayes, the value of d_N was plotted against the value of d_S for H7N1, H7N2, H7N3 and H7N7, so that the rates for different subtypes could be directly compared. It may be observed that, whilst the d_S values were similar for all four subtypes, there was little overlap between the H7N2 d_N values and those for the other subtypes. For each subtype, the linear regression line for the d_N value for a tree sample against the d_S value for the tree sample is shown.

4.5.4 Site-by-site analysis of H7 HA1 d_N/d_S on different NA subtype backgrounds

Estimates of d_N and d_S at individual H7 HA1 codon sites were calculated separately for each NA background subtype in order to investigate the process driving differences in selective pressure between H7 HA1 on an N2 NA background, compared to on an N1, N2 or N3 background, and to identify sites under putative positive selection. Of the 329 codon sites studied, the vast majority (more than 96% for all NA subtype backgrounds) had a d_N/d_S ratio of less than 1. A small number of sites were identified as being under putative positive selection, i.e. having $d_N/d_S > 1$, and such sites were distributed across the alignment length (Figure 4.7 and Table 4.4). The domain in which each site with $d_N/d_S > 1$ was observed was recorded. Sites under putative positive selection were observed in all domains: the signal peptide region, which directs the HA protein to the virion surface; the fusion domain, which fuses the HA protein to the rest of the virion; the receptor binding domain, which binds to sialic acid receptors in host cells, and the vestigial esterase domain, whose metabolic role is redundant but which has been speculated to play some part in the membrane fusion activity of modern-day influenza viruses (Sun *et al.* 2012).

The largest number of sites under putative positive selection was observed on the N2 NA background (23 sites out of the 329 considered). This was approximately twice the number of sites with $d_N/d_S > 1$ on N1, N3 or N7 backgrounds (13, 9 and 8 sites respectively) (Table 4.4). When the largest 50 d_N/d_S values across the HA1 codon sites were ordered by magnitude for each NA background subtype, the d_N/d_S value on the N2 background was higher than the d_N/d_S value of that rank on all other NA subtype backgrounds (Appendix B, Figure B2a). The large d_N/d_S values observed at individual codon sites for H7 HA1 on the N2 NA background would have led to the elevated HA1-wide d_N/d_S observed on the N2 NA background; however, H7N2 also had many of the smallest d_N/d_S values out of the different subtypes at individual codon sites (Figure 4.8 and Appendix B, Figure B2b and Figure B3). For all NA subtype backgrounds, sites with d_N/d_S >1 were observed in each of the fusion, vestigial esterase and receptor binding domains.

Although high d_N/d_S values were observed at two sites in the signal peptide region of H7 HA on NA backgrounds N2, N3 and N7, no sites with d_N/d_S >1 were observed for the H7 HA signal peptide region on the N1 NA background. The signal peptide region has previously been included in gene-wide calculations of d_N/d_S for influenza HA (e.g. by Fitch *et al*. (1991) and Chen and Holmes (2006)), and the alignment-wide d_N/d_S values across the alignment which are reported in Section 4.5.3 encompass the signal peptide region and HA1 coding region. Note that the same pattern of average d_N/d_S across sites was observed corresponding to the different NA background subtypes (i.e. a higher d_N/d_S when H7 HA was on an N2 NA background than on an N1, N3 or N7 NA background) when values were averaged over just the HA1 region (i.e. excluding the signal peptide region) as when the signal peptide region was included (Appendix B, Table B2 and Table B3). However, the difference between the d_N/d_S and d_N distributions was less extreme, particularly for the N1-N2 comparison, when the signal peptide region was excluded than when it was included.

Some commonality was observed between the H7 HA1 sites with d_N/d_S >1 on different NA subtype backgrounds. One site (site 218 in H3 numbering) had d_N/d_S >1 on all four NA subtype backgrounds; 3 sites had d_N/d_S >1 on 3 out of the four NA subtype backgrounds and 10 sites had d_N/d_S > 1 in two out of the four background NA subtypes (Table 4.4). Site 218 has been linked with receptor-binding specificity (Daniels *et al.* 1987; Connor *et al.* 1994; Skehel and Wiley 2000) and thus high levels of non-synonymous change at this site could signify a move towards viruses which are capable of infecting other host species.

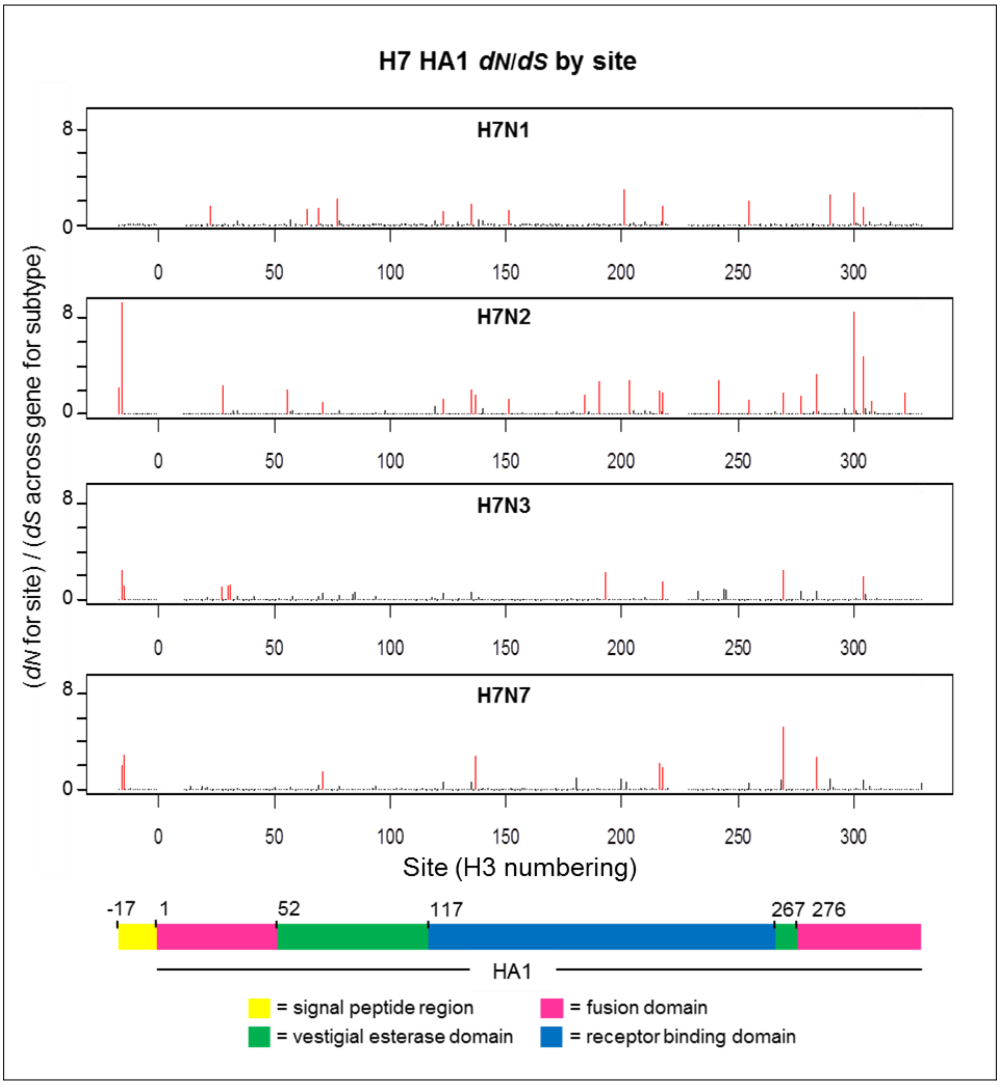


Figure 4.7
Distribution of d_N/d_S values across the avian influenza H7 HA1 subsegment, on different NA subtype backgrounds. The d_N value for each site was divided by the average d_S across all sites for that subtype to obtain a d_N/d_S value for each site on each background NA subtype. Sites with $d_N/d_S > 1$, i.e. under putative positive selection, are highlighted in red. Sites under putative positive selection were distributed across the HA1 for all background NA subtypes. Although there is some variation between NA backgrounds in terms of the sites under putative positive selection, there is also some commonality between the subtypes (see Table 4.4). A coloured key is provided, which indicates the HA1 domain: fusion (pink), vestigial esterase (green) or receptor binding (blue). The signal peptide region (17 amino acids preceding the HA1) is indicated in yellow.

H7 HA1 Site (H3 numbering)	H7N1	H7N2	H7N3	H7N7
-17		X		
-16		X	X	X
-15			X	X
22	X			
27			X	
28		X		
30			X	
31			X	
56		X		
64	X			
69	X			
71		X		X
77	X			
123	X	X		
135	X	X		
137		X		X
151	X	X		
184		X		
190		X		
193			X	
201	X			
203		X		
216		X		X
218	X	X	X	X
242		X		
255	X	X		
270		X	X	X
277		X		
284		X		X
290	X			
300	X	X		
304	X	X	X	
308		X		
322		X		

Table 4.4
H7 HA1 sites with d_N/d_S>1 in stochastic mutational analysis on different NA subtype backgrounds. As is influenza convention, sites are numbered according to the H3 numbering for HA1 (positive sites numbers) and the peptide signal region (negative site numbers). Site numbers are coloured according to the domain: fusion (pink), vestigial esterase (green) or receptor binding (blue).

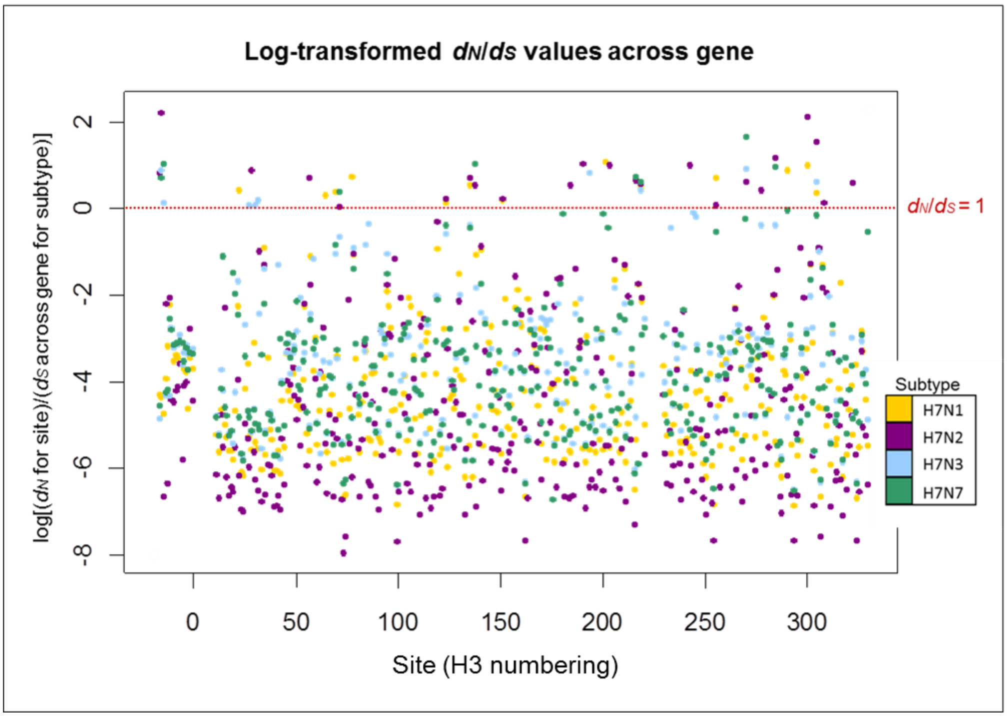


Figure 4.8

Log(d_N/d_S) values across the avian influenza H7 HA1 subsegment for each NA subtype background. The natural logarithm of the d_N/d_S values from Figure 4.7 was taken, so that sites with log(d_N/d_S) > 0 corresponded to d_N/d_S > 1, and sites with log(d_N/d_S) < 0 corresponded to d_N/d_S < 1 (the value log(d_N/d_S) = 0, i.e. d_N/d_S =1, is shown as a dotted red line). The d_N/d_S values for each site are colour coded according to the background NA subtype. Codon sites correspond to the H3 numbering.

4.5.5 Mutational mapping analysis by virus pathogenicity and avian host

The datasets for H7 HA sequences grouped by background NA subtype were not uniform in terms of their pathogenicity and host composition (Table 4.1). For example, all of the H7N2 sequences were LP, whilst approximately half of the H7N7 isolates were HP. Also, 27% of H7N7 sequences were isolated from aquatic poultry, compared to just 10% for H7N2. In order to assess the potential impact of differences in dataset composition as a confounding factor in making comparisons of evolutionary rates between subtypes, further mutational mapping analysis was conducted so that d_N, d_S and d_N/d_S could be compared for lineages corresponding to highly pathogenic (HP) and low pathogenic (LP) avian influenza viruses, and for viruses isolated from different avian hosts. Figure 4.9 shows HPD plots for d_N, d_S and d_N/d_S for HP and LP lineages. As may be observed from the HPD plots, the d_N/d_S mean and HPD intervals (Table 4.5) and the randomised pairs analysis (Table 4.6), the distributions of d_N, d_S and d_N/d_S did not differ significantly between HP and LP lineages. Likewise, there was no substantial difference in the distributions of evolutionary rates between lineages corresponding to viruses sampled from avian host orders Anseriformes or Galliformes (Figure 4.10, Table 4.7 and Table 4.8).

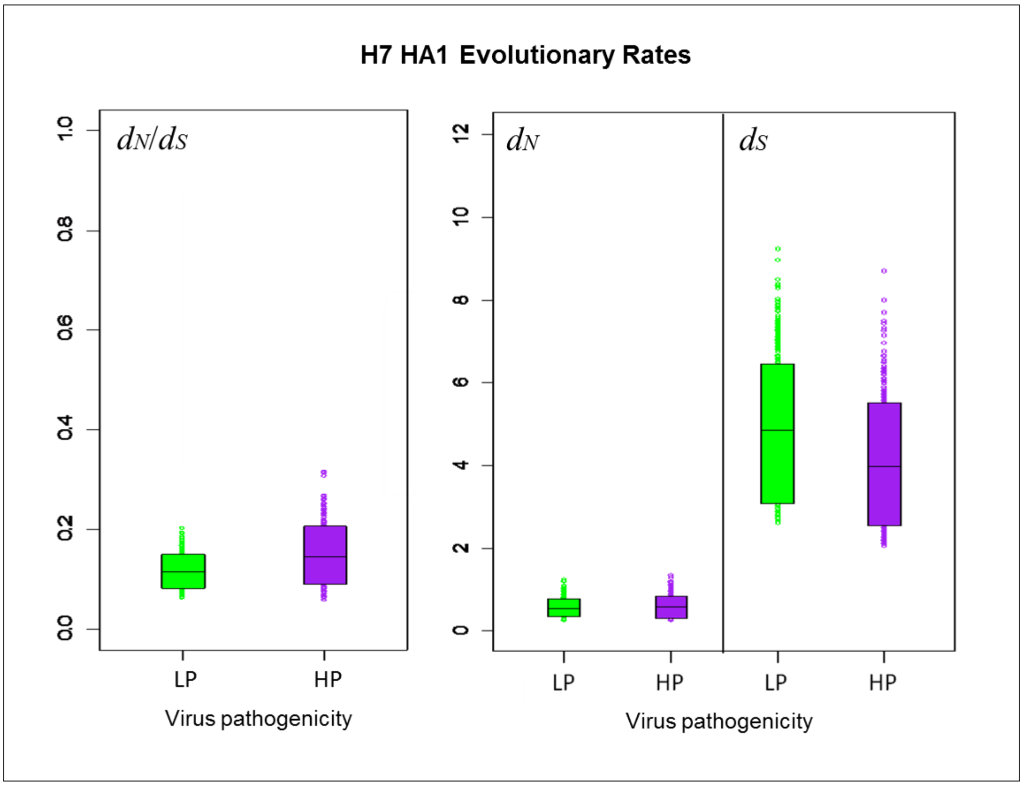


Figure 4.9
90% HPD plots for d_N/d_S, d_N and d_S for H7 HA1 along branches classified by virus pathogenicity. The coloured boxes show the limits of the narrowest interval containing 90% of the posterior estimates. The horizontal lines inside the boxes indicate the location of the mean for HP or LP viruses. The similarity in evolutionary rates for HP and LP viruses can be observed from the overlap in the distributions and the location of the means of the distribution for HP viruses within the 90% HPD limits of the corresponding LP distribution and *vice versa*.

Virus pathogenicity	Mean d_N/d_S	Lower 90% HPD limit for d_N/d_S	Upper 90% HPD limit for d_N/d_S
HP	0.146	0.092	0.207
LP	0.115	0.082	0.150

Table 4.5

Average d_N/d_S across the H7 avian influenza HA1 for viral lineages of different pathogenicities. Means and 90% HPD limits of the posterior distributions for d_N/d_S were estimated along parts of the phylogeny samples corresponding to HP and LP viruses. The average d_N/d_S across the HA1 coding region for HP and LP lineages was obtained for each MCMC tree sample by first averaging over mutational mapping replicates on that tree, then calculating average values for d_N and d_S across all HA1 sites. The site-averaged d_N was then divided by the site-averaged d_S for the part of the tree corresponding to HP viruses (and similarly for LP viruses) to obtain 1000 posterior estimates of the d_N/d_S ratio.

Comparison	d_N/d_S	d_N	d_S
HP-LP	0.763821 0.236179	0.519682 0.480318	0.26037 0.73963

Table 4.6

Randomised pairing analysis for H7 HA1 mutational mapping analysis split by pathogenicity. Evolutionary rate estimates were compared for highly pathogenic (HP) and low pathogenic (LP) lineages. The proportion of randomised pairings for which the LP value subtracted from the HP value was positive (top value in each cell) and negative (bottom value in each cell) was reported. Under the paired differences analysis, the posterior distributions of rates did not appear to differ significantly between HP and LP viruses, since an extreme split between the number of positive and negative differences was not observed.

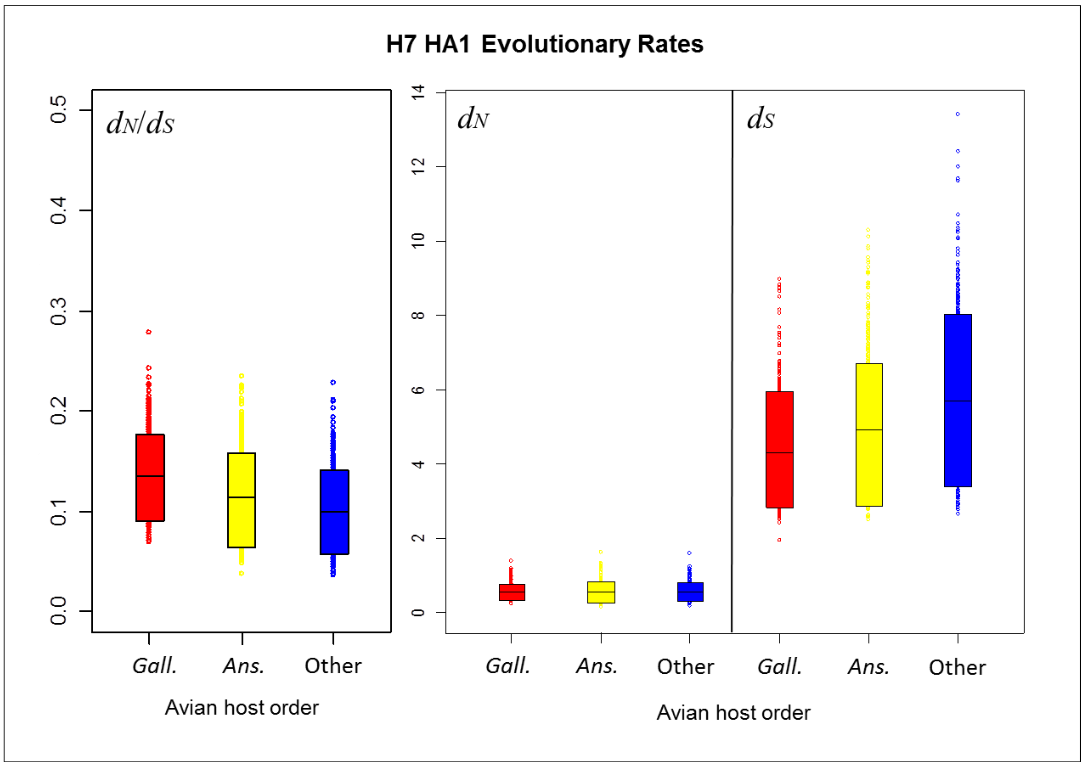


Figure 4.10

90% HPD plots for d_N/d_S, d_N and d_S for H7 HA1, split by avian host order. The means and HPD limits for d_N/d_S and rates of synonymous and non-synonymous substitution were similar for anseriform, galliform and other avian hosts. This indicated that the taxonomic order of the avian host from which influenza viruses were isolated did not have a significant effect on evolutionary rates or selective pressure experienced by the virus.

Avian host order	Mean d_N/d_S	Lower 90% HPD limit for d_N/d_S	Upper 90% HPD limit for d_N/d_S
Anseriformes	0.113	0.065	0.158
Galliformes	0.135	0.091	0.177
Other	0.100	0.057	0.141

Table 4.7

Average d_N/d_S across the H7 avian influenza HA1 for viral lineages corresponding to different avian host orders. Stochastic mutational mapping was used to calculate the d_N/d_S along lineages corresponding to viruses from different avian host orders (Anseriformes, Galliformes and others) for 1000 MCMC tree samples, in an analogous manner to that described for comparisons by background NA subtype and viruses of different pathogenicity. The means and HPD limits corresponding to different host orders were compared and indicate that was no substantial difference in average selective pressure across the HA1 between viruses in different avian host orders.

Comparison	d_N/d_S	d_N	d_S
Ans. - Gall.	0.293355	0.443505	0.647044
	0.706645	0.556495	0.352956
Ans. - other	0.637318	0.482577	0.336128
	0.362682	0.517423	0.663872
Gall. - other	0.821002	0.541115	0.213498
	0.178998	0.458885	0.786502

Table 4.8

Randomised pairing analysis for H7 HA1 mutational mapping analysis split by avian host order. Evolutionary rates were calculated along lineages corresponding to viruses from different avian hosts, and rate distributions corresponding to lineages from Anseriformes (Ans.) and Galliformes (Gal.) were compared. The proportion of randomised pairings for which the difference between rate estimates for the first and second taxonomic order listed was greater than 0 (top value in each cell) and less than 0 (bottom value in each cell) was reported in the table. The randomised pairing analysis indicated that the posterior distributions of rates did not differ substantially between anseriform and galliform hosts, or other avian host orders.

4.6 Discussion

4.6.1 Study design

Reassortment repeatedly exposes influenza segments to different genetic backgrounds. The aim of this study was to investigate how evolutionary rates of H7 avian influenza HA, and the selective pressure it experiences, are affected by the background NA subtype of the virus. Comparing the evolution of H7 HA on separate datasets corresponding to different NA subtype backgrounds (as was performed by Chen and Holmes 2006) would have made the implicit assumption that sequences cluster perfectly according to the NA subtype of the virus in a tree of all H7 HA sequences. The observed violation of this assumption has the potential to introduce error into estimates of evolutionary rates because evolution along ancestral lineages linking sequences of the same subtype from different parts of the tree would contribute to the rate estimate.

The stochastic mutational mapping method employed in this chapter enabled d_N and d_S to be calculated along parts of the HA tree corresponding to different NA subtype backgrounds, so that H7 HA sequences from viruses of various NA subtypes could be analysed concomitantly without making prior assumptions about their evolutionary relationships. The d_N/d_S ratio could be calculated simply by summarising the relevant lineages (to which the same NA subtype had been assigned), and did not require additional parameters to be introduced into the model. The rescalings also allowed d_N and d_S to be compared between clades of different sizes and divergence. Overall, the results indicated that d_N/d_S was higher for H7 HA1 on an N2 background than on an N1, N3 or N7 background. Methods such as SLAC, REL, FEL (described in Chapter 2, Section 2.11) produce d_N/d_S estimates across the whole tree, so would only be applicable to individual clades of H7 HA corresponding to a particular NA background subtype for this study. The mutational mapping method also provides more detailed evolutionary information than SLAC, REL and FEL, which do not provide detailed information about the timings and nature of changes along the tree at each site. Although the DEPS method of

Kosakovsky Pond *et al*. (2008) can detect directional evolution towards particular residues, again this does not provide all of the information required for the rescalings used in this chapter.

Chen and Holmes (2006) used single likelihood ancestor counting (SLAC) (Kosakovsky Pond and Frost 2005) to obtain point-estimates for d_N/d_S based upon a single neighbor-joining tree for datasets corresponding to each background NA subtype. They found the H7N2 d_N/d_S estimate to be the highest (0.17, compared with 0.11, 0.09 and 0.09 for H7N1, H7N3 and H7N7 respectively) (Chen and Holmes (2006), Supplementary Table), consistent with those obtained in the more evolutionarily and statistically robust stochastic mutational mapping study described in this chapter. For each subtype, the SLAC point estimate of d_N/d_S obtained by Chen and Holmes fell within the 90% HPD limits from the mutational mapping analysis, despite the use of non-identical sets of sequences, different study designs and different methods of inference. The error introduced by creating individual datasets corresponding to different NA backgrounds therefore did not appear to be large in this case. This is perhaps because the amount of evolutionary time along the deeper part of the tree (in which error could be introduced in an analysis of separate datasets for each NA subtype) was small compared to the amount of evolutionary time encompassed by correctly labelled clades or lineages towards the tips of the tree.

4.6.2 Advantages of stochastic mutational mapping for calculating d_N/d_S

The Bayesian stochastic mutational mapping method employed in this chapter possesses many advantages for investigating selective pressure in avian influenza H7 HA on different NA subtype backgrounds, in the presence of reassortment. Bayesian methods for phylogenetic inference and mutational mapping provide an advantage over parsimony and maximum-likelihood methods since they naturally accommodate uncertainty in the phylogenetic reconstruction (by considering multiple tree and substitution model samples) and the mutational history (by sampling multiple histories for each site in each phylogeny sample). Failing to account for phylogenetic uncertainty can lead to artificially narrow confidence intervals when estimating substitution rates (Nielsen 2002). Both Bayesian and ML methods share the advantage over parsimony that non-parsimonious maps are not automatically excluded. Using parsimony to minimise the number of mutations required to produce the observed pattern in the data can lead substitution rates to be underestimated, perhaps by a factor of over 20% for influenza (Nielsen 2002). Parsimony inference can also bias d_N/d_S estimates by underestimating the number of synonymous changes in scenarios where synonymous mutations occur more frequently than non-synonymous mutations (Nielsen 2002).

In addition to the ability to sample multiple mutational histories, the mutational mapping method possesses other advantages over the PAML maximum likelihood software (Yang 1997; Yang 2007). Although PAML can be used to estimate d_N/d_S along the branches of a phylogeny (Yang 1998; Yang and Nielsen 1998), in this study there would have been a lack of power for statistical testing for positive selection using likelihood-ratio tests, since parameters are estimated for each branch in the tree (approximately 500 branches for the H7 HA dataset). In addition, whilst the stochastic mutational mapping approach allows d_N/d_S to be estimated for each site along all individual branches of the tree, the branch-site models in PAML require branches with potentially positively selected sites to be pre-specified for testing. Mutational mapping also records the timings of mutations across the tree in the sampled mutational map, which can be used in calculating evolutionary rates, whereas existing maximum likelihood methods do not.

By performing the rescalings described in Section 4.4.4, it was possible to extend existing mutational mapping methods to estimate rates of synonymous substitution (d_S) and non-synonymous substitution (d_N), rather than merely counting the number of synonymous or non-synonymous changes along branches (Nielsen 2001; Bollback 2006). Also, estimating d_N and d_S separately allowed differences in the d_N/d_S ratio to be attributed to underlying differences in the non-synonymous or synonymous rate, so that population genetic explanations for such differences could be considered (see Section 4.6.3). Note that future studies could use similar rescalings in conjunction with recently developed robust counting methods (O'Brien *et al.* 2009) to estimate d_N and d_S by analytically solving for the number of synonymous and non-synonymous changes rather than using the more time-consuming stochastic simulation methods.

4.6.3 Evolutionary implications

Mutational mapping analysis indicated that H7 avian influenza HA1 from viruses with an N2 NA subtype was subjected to substantially different selective pressures than when associated with an N1, N3 or N7 NA background. The elevated d_N/d_S ratio for H7N2 lineages was attributed to an increase in the rate of non-synonymous substitution, whilst synonymous substitution rates remained constant across subtypes. HA1-wide estimates of d_N/d_S were less than one for all background NA subtypes, consistent with earlier work which has suggested that the influenza HA is conserved overall (e.g. Fitch *et al*. (1991), Ina and Gojobori (1994), Bush *et al*. (1999)).

Rates of synonymous and non-synonymous change can be affected by population genetic factors. Under selective neutrality, the rate of substitution is independent of the effective size of the population and is equal to the mutation rate (Kimura 1968). Assuming that all synonymous changes are essentially neutral, d_S is independent of the effective population size N_e and is simply the mutation rate. The observation in this study that the posterior rate estimates of d_S for H7 HA1 were constant across different NA subtype backgrounds suggested that the mutation rate was constant across subtypes. There is no *a priori* reason why the mutation rate of H7 avian influenza HA1 should be affected by the NA subtype of the virus. Rather, the mutation frequency for both the HA and NA segments is most likely to be influenced by changes in the genes involved in viral replication (Holland *et al.* 1982), for example those affecting the error-proneness of the polymerases. Previous researchers have found synonymous rates to be generally more uniform than corresponding non-synonymous rates in other organisms, for example across mammalian genes encoding different proteins (Miyata 1984; Kimura 1986).

Under non-neutral models of evolution, differences in selective pressure could lead to differences between substitution rates (Kimura 1968). Since non-synonymous changes in the HA1 coding region are likely to be non-neutral, the elevated d_N observed for avian influenza H7 HA1 on an N2 NA subtype background, compared to on N1, N3 and N7 backgrounds, might be explained by a number of scenarios. Firstly, selection could be acting to fine-tune the functional HA-NA balance of H7 HA on an N2 NA background following reassortment. Secondly, a burst of positive selection could have occurred in the H7N2 lineages, which is not a consequence of the N2 NA background, but instead a consequence of an unrelated, covarying factor such as avian host, demographic scenario, or an interaction with another gene segment. Thirdly, a relaxation of selective constraint could have taken place when H7 HA was exposed to the N2 NA background. The results of this study do not definitively distinguish between such scenarios and causality cannot be inferred. However, whilst d_N/d_S >1was observed in a larger number of HA1 sites on the N2 NA background than on N1, N3 or N7 backgrounds, at many sites the N2 viruses also had the lowest d_N/d_S values out of all NA subtype backgrounds and this is not indicative of an overall relaxation of selective constraint. One explanation for the observed pattern of site-by-site d_N/d_S values could be a larger effective population size in HA for the H7N2 viruses, which would allow selection to act more effectively in removing deleterious mutations, leading to a reduction of variation at some sites. Another explanation could be a selective sweep which has resulted in a reduction of background variation in H7N2 HA.

4.6.4 H7N2 avian influenza

Of the 75 H7N2 HA1 sequences studied in this chapter, 66 were from viruses circulating in the North American live bird markets between 1994 and 2006, or from the many avian influenza outbreaks they seeded in commercial poultry in the Northeast United States during this period (Senne *et al.* 2003a; Spackman *et al.* 2003). Except for one isolate from 1999, which clustered with contemporaneous North American H7N3 sequences, the North American H7N2 isolates formed a monophyletic clade in the HA1 MrBayes consensus tree (posterior probability 0.98). The nine Eurasian H7N2 isolates covered a large geographic area (Hong Kong - UK) and timespan (1978-2007) and did not cluster together, instead being distributed across the Eurasian clade.

It may also be noted that 88% of the North American H7N2 sequences possessed a deletion of 8 amino acids at the HA receptor binding site, and a recent study has put forward the idea that non-synonymous changes might have occurred in the HA to maintain functionality (Yang *et al.* 2010). This could be compatible with the observation that a large number of sites with d_N/d_S >1 in this study were found in the receptor binding domain for H7 HA on the N2 NA background. It is possible that molecular changes at, or adjacent to, other sites in the receptor binding region (for example, the elevated d_N/d_S observed in H7N2 at HA1 sites 216 and 218) could be compensating for the 8 amino acid receptor binding site deletion. Although this could indicate co-evolution within the HA, again this could be to restore HA activity levels to match those of the N2 NA. Future studies might investigate the evolution of lineages associated with the receptor binding site deletion further by determining whether elevated levels of non-synonymous change only applied to these lineages.

Mutational mapping analysis indicated that non-uniform dataset composition with respect to virus pathogenicity or proportion of viruses from anseriform or galliform hosts was not responsible for the elevated d_N/d_S observed in H7N2 avian influenza HA1, since no differences in d_N/d_S were observed between these groups. However, other differences between the environments from which sequences were isolated may have influenced the selective pressure experienced. It has been suggested that long term evolution in commercial poultry, which are not the natural reservoir of avian influenza, could lead to accelerated rates of evolution and the accumulation of point mutations in viruses in the live bird markets (Webster 1998; Senne *et al.* 2003b). N2 was the only background NA subtype for which a number of H7 HA sequences from the live bird markets were available, thus it would have been difficult to separate the effect of subtype and demography in testing for differences in d_N/d_S. However, 41 out of 62 (66%) of the H7N1 sequences were sampled during an outbreak of LP and

HP H7N1 avian influenza in poultry in Italy, and the elevated d_N/d_S did not extend to this subtype (although sequences were sampled over a period of less than two years, compared to over 12 years for H7N2 in the North American live bird markets).

H7N2 was the most common avian influenza subtype isolated from North American live bird markets between 1994 and the mid-2000s (Panigrahy *et al.* 2002; Suarez *et al.* 2003), garnering attention as a potential source for a human pandemic virus (Belser *et al.* 2009). Indeed, a pre-pandemic vaccine based upon the North American H7N2 HA has already been developed (Pappas *et al.* 2007). North American H7N2 viruses proved capable of causing human infection: respiratory illness and neutralising antibodies were reported for a worker in the live bird markets (CDC 2004b) and hospital treatment was required for an immunocompromised patient from whom the H7N2 virus was isolated (CDC 2004a). Although human infection with H7N2 avian influenza is rare, viruses isolated in North America between 2002 and 2003 have been found to exhibit increased affinity towards human-like α-2,6-linked sialic acid receptors (Belser *et al.* 2008). This could signify an increased risk of an H7N2 influenza virus becoming transmissible amongst human hosts. H7N2 has so far only presented in a low pathogenic form, but molecular evidence suggests an accumulation of basic amino acids at the HA cleavage site which could result in the emergence of a highly pathogenic virus (Spackman *et al.* 2003). The elevated d_{N} for H7N2 avian influenza observed in this study may indicate a heightened risk of molecular changes occurring which could increase the pathogenicity of the virus, or its ability to infect new host species and become transmissible amongst humans.

4.6.5 Future directions

Future studies could investigate variation in d_N/d_S between different branches to which the same NA background subtype had been assigned, for example to determine whether the high level of non-synonymous change on the N2 NA background occurred only amongst the large clade of North American H7N2 lineages, or whether it extended to the Eurasian H7N2 lineages. In addition, d_N/d_S

estimates could be obtained for branches immediately following an event where the NA background subtype had changed (i.e. following a reassortment event), or after introduction to a new avian host species (Shackelton *et al.* 2005), to determine whether these branches exhibited an elevated rate of adaptive evolution. The effect of continuous virus circulation in non-natural avian hosts in the live bird markets upon d_N/d_S could also be investigated in other datasets. For example, further studies could compare the selective signatures of North American H7N2 with H5N1 viruses which are endemic in the live bird markets of East Asia (Li *et al.* 2004), or could compare d_N/d_S along lineages corresponding to wild or domestic hosts. Furthermore, evolutionary rates along terminal and internal branches could be compared in future studies (or d_N/d_S could be compared on different NA subtype backgrounds using only internal branches) since elevated d_N/d_S ratios have previously been observed towards the tips of phylogenies (Sharp *et al.* 2001).

The results of this study are consistent with the hypothesis that reassortment exposes HA to significant changes in selective forces via association with different NA subtypes. Although it is not possible to establish a causal relationship between background NA subtype and differences in evolutionary rates of HA, one way in which more detailed investigations could be carried out is through the use of stochastic character trait mapping (Huelsenbeck *et al.* 2003). Multiple traits could be mapped onto the trees simultaneously (e.g. NA subtype and host), which would allow correlations between traits to be formally assessed and would enable d_N and d_S to be calculated for different trait combinations (e.g. H7N7 from Anseriformes and H7N7 from Galliformes; H7N2 from North America and H7N2 from Eurasia).

In this study, d_N/d_S was found to be elevated for H7 HA on an N2 background, compared to on N1, N3 or N7 NA backgrounds. The HA protein, and in particular the part encoded by the HA1 subsegment, is of major antigenic significance for the influenza virus. Along with NA and NS1, HA has been found to exhibit a higher d_N/d_S than other genes (Chen and Holmes 2006). It is possible that, due to its close interaction with HA, the evolution of NA could be driven by the HA segment, so that viruses with a quickly evolving HA also have a fast-evolving NA. Future analysis could investigate whether the elevated d_N/d_S observed for H7N2 HA1 extends to the corresponding NA sequences for those viruses, compared to the NAs for the other subtypes. Interactions with other influenza proteins, such as the matrix protein, with which the HA and NA both interact closely, could also be investigated.

Previous researchers have tended to focus on human seasonal influenza (e.g. Bush *et al*. (1999), Suzuki (2006)), or viruses which have undergone transmission or adaptation to humans (e.g. Kongchanagul *et al*. (2008)), rather than sites under positive selection in avian influenza. However, positive selection has been detected at particular sites for avian H5N1 sequences, especially in HA and NS1 (Kosakovsky Pond *et al.* 2008). Although H7 avian influenza has been less studied than H5N1, limited evidence has been provided for positive selection at HA sites using the REL and FEL methods (Lebarbenchon and Stallknecht 2011). In this study H7 HA sites under putative positive selection were identified for each of the NA background subtypes considered, based on an observed d_N/d_S value (the mean averaged over the 1000 MrBayes samples) of greater than one. However, under selective neutrality a distribution of d_N/d_S values would still be expected, some of which would be greater than one. In order to obtain statistical support for sites being under positive selection, posterior predictive p-values could be generated, which describe the location of the observed statistic within a predictive (null) distribution and may be interpreted similarly to traditional frequentist p-values (Nielsen and Huelsenbeck 2002). Multiple testing corrections, such as the false discovery rate (Benjamini and Hochberg 1995), may also be applied when large numbers of codons are being considered. HPD intervals for d_N/d_S for each site could also be considered, rather than the mean values over the MrBayes samples. For example, one could consider only sites for which the lower 95% HPD limit exceeded 1 to be under putative positive selection; this would be a more stringent measure than using the mean d_N/d_S value for a site. In the strictest case, one could generate posterior predictive p-values (see above) for the value of the lower 95% HPD limit of such sites and only consider statistical support for a site being under positive selection to have been attained when the p-value for the lower 95% HPD limit was significant (e.g. $p<0.05$).

After identifying H7 HA sites with statistical support for being under positive selection on a particular NA subtype background, the location of such sites upon the 3D structure could be determined and visualised using software such as PyMol (www.pymol.org/). The DEPS software (Kosakovsky Pond *et al.* 2008) could be used to look for evidence of directional selection towards particular residues at sites with $d_N/d_S>1$ in specific clades, in particular the North American H7N2 clade, and it is possible that the mutational mapping method could be adapted in to infer more detailed evolutionary behaviour in the future, for example to investigate variation in d_N/d_S over time.

Finally, laboratory studies could be conducted to further investigate HA-NA interactions, and their results could be integrated with those from *in silico* studies such are reported in this chapter. For example, the activity levels of different HA and NA strains and subtypes could be determined, to consider whether there is evidence for post-reassortment adaptation amongst the strains considered in this study. It may also be fruitful to investigate whether there are systematic differences between receptor binding activity levels of particular HA or NA subtypes. In addition, the precise nature of the genetic changes which take place when HA is placed in a novel NA background (or *vice versa*) could also be explored in the laboratory using reverse genetics experiments, to provide an insight into how the balance between HA and NA activity is regulated.

Chapter 5

Mapping discrete trait evolution on avian influenza phylogenies

5 Mapping discrete trait evolution on avian influenza phylogenies

5.1 Chapter Summary

In Chapter 3, the distribution of viral NA subtypes across the tips of avian H7 HA phylogenies provided visual evidence for reassortment of different NA subtypes onto the H7 HA background. The distributions of wild and domestic avian hosts, and viruses of low and high pathogenicity (LP and HP respectively), were also visualised at the tips of consensus phylogenies. In Chapter 4, a parsimony algorithm was used to assign viral NA subtypes along ancestral branches, although the underlying transition process was not quantified. In this chapter, I apply state-of-the-art discrete trait mapping methods in BEAST, which were originally developed for phylogeographic inference, to quantitatively analyse the distribution of discrete traits upon time-scaled avian H7 HA phylogenies. I consider the distribution of viral NA subtypes across the trees, which is shaped by reassortment, and investigate whether there is evidence for higher rates of transition between particular subtypes. I also analyse patterns of avian host transition (from wild to domestic birds, or *vice versa*) and viral pathogenicity (HP or LP) across the avian H7 HA phylogenies. The extent to which discrete trait mapping methods can be used to test evolutionary hypotheses about avian H7 HA evolution is limited by the available data. However, with sufficient sampling these techniques could prove a powerful framework for quantifying trait evolution across phylogenies, as I shall demonstrate in an investigation of inter-subtype recombination in HIV-1 group M in Chapter 6. I also conduct a technical investigation of the use of different models and procedures for discrete trait mapping analyses in BEAST and discuss their relative merits and applicability in different scenarios.

5.2 Chapter Aims

- Investigate the use of methods in BEAST for mapping non-geographic discrete traits onto phylogenies, in particular the subtype of another viral protein (here, NA subtype on an H7 HA phylogeny), avian host type (wild or domestic) and pathogenicity (LP or HP).
- Use discrete trait mapping methods in BEAST to quantify how often H7 HA is exposed to different NA subtype backgrounds across the phylogeny as a result of reassortment, and determine whether transition rates are higher between particular pairs of NA subtypes.
- Identify directionality in transitions between avian hosts, or in jumps between viral pathogenicity, i.e. determine whether there is evidence for higher rates of transition from wild to domestic birds than from domestic to wild birds and from LP to HP rather than from HP to LP.
- Investigate the applicability and robustness of implementations in BEAST for testing evolutionary hypotheses about discrete trait transition across viral phylogenies.

5.3 Introduction

Avian influenza virus sequences can be labelled by properties of the virus itself, as well as the time, place and host from which it was sampled. For example, avian influenza viruses may be labelled according to whether their avian host was wild or domestic, whether the virus was of high or low pathogenicity and the background NA subtype of the virus. As demonstrated for H7 avian influenza HA sequences in Chapter 3, the distribution of discrete trait states for avian host, pathogenicity and NA subtype may be visualised on a phylogeny by colouring the tips or external branches of a tree. The observed distribution of discrete trait states at the tips of the phylogeny results jointly from character state transitions and nucleotide substitution.

Discrete trait mapping methods such as those described in Section 2.10 allow ancestral character states to be inferred across phylogenetic trees. Not only may the ancestral states be visualised along the branches, but also the dissemination of character states across the tree can be quantified. This allows hypotheses about the distribution of trait states across the tree to be tested in a formal manner. One of the most common applications of discrete trait mapping methods in infectious disease phylogenetics has been in phylogeographic inference, where information about the location (e.g. country or city) from which sequences were sampled is mapped onto phylogenies. In particular, recent implementations in the BEAST software have allowed the spatio-temporal spread of viral epidemics to be examined on the same time-scale as the accumulation of genetic diversity for fast evolving pathogens (Lemey *et al.* 2009). Methods for mapping discrete traits upon phylogenies have also been used to study disease transmission between different host species (Goldberg 2003; Chen and Holmes 2009; Weinert *et al.* 2012), and to investigate temporal clustering and geographic subdivision of influenza viruses in wild birds (Chen and Holmes 2009).

In Chapter 3 and Chapter 4, the distribution of viral NA subtypes across the tips of avian H7 HA phylogenies as a result of reassortment was observed. Traditionally, reassortment has been detected between influenza RNA segments by looking for phylogenetic discordance between trees constructed from different segments for the same set of taxa (e.g. Holmes *et al*. (2005), Macken *et al*. (2006), Nelson *et al*. (2008), Vijaykrishna *et al*. (2008a)), or in other viruses by comparing phylogenies for different regions of the genome (e.g. Robertson *et al*. (1995)). Although algorithms have been developed to detect reassortment by comparing topologies from different influenza gene segments using graph-mining (for example, the GiRaF software of Nagarajan and Kingsford (2011)), such methods only detect incongruent taxa and do not provide a measure of the discordance between phylogenies which allows the amount of recombination to be quantified.

Reassortment which creates different combinations of influenza HA and NA subtypes can naturally be investigated using discrete trait mapping because the subtypes themselves can be mapped onto the phylogenies. In this chapter I use recently-developed Bayesian discrete trait mapping methods in BEAST (Minin and Suchard 2008a; Minin and Suchard 2008b; Lemey *et al.* 2009; O'Brien *et al.* 2009; Talbi *et al.* 2009) to investigate reassortment of Eurasian avian H7 HA with viruses of different NA subtypes. Such methods can be used to obtain estimates of the overall rate of trait transition across time-scaled phylogeny samples (e.g. NA subtype transition on the H7 HA phylogeny, as a result of reassortment) and to calculate relative rates and numbers of transitions between particular subtypes. Transitions between wild and domestic avian hosts, and between viruses of low and high pathogenicity, are also considered, allowing the processes observed in Chapter 3 and Chapter 4 to be quantified. I consider technical aspects of the discrete trait mapping methods to determine the most appropriate manner in which to use them to investigate reassortment or recombination, and identify a dataset in which a more detailed and comprehensive analysis of joint HA and NA influenza subtypes could be performed in the future.

5.4 Methods

5.4.1 Dataset composition

In Chapter 3 and Chapter 4, phylogenetic analysis revealed that avian H7 HA sequences clustered into two major clades: one corresponding to sequences from Eurasia, Africa and Australasia and the other corresponding to sequences from North and South America. For the discrete ancestral trait mapping analyses in this chapter, a dataset consisting of Eurasian sequences was selected rather than using a global avian H7 HA dataset; this dataset was a subset of the sequences analysed in Chapter 3 (Appendix C, Table C1). The existence of distinct avian H7 HA populations in Eurasia and North America, and the observed difference in the distribution of background NA subtypes in these regions, suggested that each might require a different discrete trait transition model (Appendix C, Figure C1). Eurasia was chosen as the focus of this study rather than North America, due to better sampling of data; in particular, in Chapter 3 and Chapter 4 there appeared to be a sampling bias in North America towards LP H7N2 sequences from domestic birds. Australasian sequences were not included in the Eurasian sequence dataset in this chapter due to the suggestion of the maintenance of a distinct Australasian H7 HA population (Chapter 3; Bulach *et al*. (2010)). Only sequences sampled after 1990 were included because, in the analysis presented in Chapter 3, many sequences from before 1990 did not lie in the expected position in a plot of root-to-tip divergence. The dataset used in this chapter, henceforth referred to as the 'Eurasian post-1990 dataset', consisted of 159 sequences for which the viral NA subtype, the pathogenicity of the virus and the avian host status (wild or domestic) were known. Subtypes N1, N3 and N7 each accounted for approximately one third of the dataset, whilst there were less than 6 sequences for each of the N2, N8 and N9 subtype backgrounds (Table 5.1). 37% of the sequences were from wild avian hosts and 23% were from highly pathogenic viruses.

Trait	State	Number of sequences	Frequency
Subtype	H7N1	56	0.352
	H7N2	6	0.038
	H7N3	44	0.277
	H7N7	46	0.289
	H7N8	3	0.019
	H7N9	4	0.025
Host	wild	59	0.371
	domestic	100	0.629
Pathogenicity	LP	122	0.767
	HP	37	0.233

Table 5.1
Breakdown of Eurasian post-1990 avian H7 HA sequence dataset. A total of 159 Eurasian sequences were analysed, all of which were sampled after 1990. Sequences were labelled according to the NA subtype of the virus, whether the avian host was wild or domestic and whether the virus was of low or high pathogenicity (LP or HP respectively), as described for the larger avian H7 HA dataset in Chapter 3. Frequencies were also calculated for each state, to enable the relationship between reported transition rates and sample sizes to be assessed.

5.4.2 Discrete ancestral trait mapping analysis

The dissemination of viral NA subtypes, avian host types (wild or domestic) and HP and LP viruses across the avian H7 HA phylogenies was investigated using discrete trait mapping methods in BEAST. Two discrete trait mapping approaches were taken: firstly, using the methods developed and implemented in BEAST for discrete phylogeography by Lemey *et al*. (2009) and secondly using the ‘Markov jumps’ methods for counting discrete trait transitions across phylogenies (described in Section 2.10.3) (Minin and Suchard 2008a; Minin and Suchard 2008b; O'Brien *et al.* 2009; Talbi *et al.* 2009). Both methods model discrete trait transition across the tree as a continuous-time Markov chain (CTMC), and both symmetric and asymmetric (reversible and non-reversible, respectively) discrete trait transition models were implemented in this chapter. In BEAST, the Markov jumps implementation uses the CTMC from the discrete trait mapping implementation of Lemey *et al*. (2009).

The output from the discrete trait mapping implementation of Lemey *et al*. (2009) (see Section 2.10.3) can be used to construct a matrix, $\mathbf{\Lambda}$, which is analogous to the **Q** matrix described in Section 2.5.1 for nucleotide substitution models. A mean instantaneous transition rate, μ, is reported, which scales the transition rates to the same time units as the tree (e.g. ‘years’ if the branch lengths of the tree are measured in years). Relative rate parameters s_{ij} are components of a matrix, **S**, and report the rate at which transition from one particular state to another occurs with respect to transitions between other pairs of states. The matrix $\mathbf{\Lambda}$ can be obtained by multiplying μ, **S** and a diagonal matrix, $\mathbf{\Pi}$, containing the equilibrium frequencies of the states (which may be estimated from the state frequencies at the tips of the tree). By exponentiating $\mathbf{\Lambda}$, the finite-time transition probabilities may be obtained.

Bayesian stochastic search variable selection (BSSVS), as implemented for discrete phylogeographic analysis by Lemey *et al*. (2009), was used when mapping background NA subtypes onto the avian H7 HA phylogenies. BSSVS aims to build a parsimonious model of the dissemination of discrete trait changes across the phylogeny, and its use in phylogeographic analyses was advocated by Lemey *et al.*

(2009) for reducing the complexity of models when a large number of distinct trait states were present at the tips of a phylogeny. A total of *n* distinct trait states in a dataset yields *n*(*n*-1) transition rates between pairs of states for a fully specified asymmetric model. Therefore, the potential number of rates in a model rapidly becomes large as the number of distinct states increases. However, there may be few or no transitions between many pairs of states across the phylogeny, and the relative rates of transition between pairs of states are difficult to inform. Such difficulties arise because the trait transition model is constructed from just one column of information (one observation for each sequence), rather than across the length of an alignment as would be the case for inference of nucleotide substitution models.

In a BSSVS analysis, individual transitions between pairs of states are switched on (indicator value = 1) and off (indicator value = 0) at different points in the MCMC. This focus on sparse matrices is due to the aforementioned difficulties in informing the relative transition rates between pairs of states. A prior distribution can be chosen to minimise the number of rates which are switched on at any step in the MCMC chain, and the proportion of the time that a rate is switched on across the chain is considered. A Bayes factor test can be used to determine the extent to which a particular rate should be included in the diffusion model (i.e. whether it is 'significantly non-zero'). The value of the Bayes factor for a particular rate being non-zero is given by the posterior odds that the rate is non-zero divided by the prior odds for the rate being non-zero, under a truncated Poisson prior with a mean of log(2) (Lemey *et al*. 2009). For the mapping of background NA subtype on the H7 HA phylogenies, there were a total of 6 distinct states (N1, N2, N3, N7, N8 and N8). The required indicator cut-offs (the proportion of the time the rate needed to be switched on across the MCMC chain) for rates to be included in the diffusion model with a Bayes factor of 3 (0.413 for the asymmetric transition model) were calculated by running the 'RateIndicatorBF', which is part of the BEAST distribution. BSSVS was also implemented for the host and pathogenicity analyses.

Whilst a non-BSSVS analysis reports relative pairwise transition rates directly, the BEAST output from a BSSVS analysis requires additional processing to account for the potential for different rates to be switched on or off across the generations, because rates are still sampled (from the prior) even when the BSSVS indicator is switched off. The indicator (1 or 0) must be multiplied by the relative rate at each sampled generation (http://beast.bio.ed.ac.uk/Discrete_Phylogeographic_Analysis), which has the effect of setting the rates which are switched off to zero at that point in the MCMC and leaving the rates which are switched on at that point unchanged.

Relative pairwise transition rate parameters from a BSSVS analysis may be averaged over the MCMC in one of two ways. Firstly, the relative rate parameters may be averaged only across the MCMC states where the rate is actually switched on (i.e. by dividing the sum of the non-zero relative rate parameters across the MCMC for a pair of states by the number of non-zero indicators for that pair of states), as was the approach taken by Lycett *et al*. (2012) for studying swine influenza reassortment using discrete trait mapping. Alternatively, the non-zero rates between a particular pair of states (and in a particular direction, if an asymmetric transition model has been implemented) may be added together and divided by the total number of MCMC samples. Averaging in this way provides a measure of the mean value of the relative rate across the MCMC, including the times when it is switched off and thus set to zero.

The output of experiments where BSSVS was implemented was compared to those where BSSVS was not used, and also to the results of Markov jumps analyses for counting labelled state changes across phylogeny samples. The Markov jumps analyses were implemented by manually editing the BEAST xml file to log the number and direction of discrete trait transitions between all pairs of states. To check for consistency, the number of Markov jumps across the tree samples was compared to the expected number of transitions under the corresponding overall mean transition rate from the CTMC. For each MCMC sample, the overall transition rate for a trait (e.g. NA subtype) from the CTMC was multiplied by the total length of the tree (the sum of the branch lengths) to obtain the expected number of changes across the tree under the inferred substitution model.

5.4.3 Generating empirical tree distributions

Discrete trait mapping in BEAST can be carried out by simultaneously constructing topologies and performing the discrete trait analysis, in which case the discrete trait information may affect the clustering of sequences in the tree by acting as another alignment column. Alternatively, a sample of phylogenies may first be inferred using the nucleotide alignment only, and the discrete trait mapping may be subsequently performed upon this 'empirical sample of trees' by manually editing the xml file (e.g. Raghwani *et al*. (2011), Lycett *et al*. (2012)). Performing the analysis in two stages means that the discrete trait information will not have been considered when inferring the tree structure. Using an empirical tree distribution has the considerable advantage of dramatically reducing computation times, since different discrete trait mapping analyses may be conducted on the same set of phylogenies. In this chapter, an empirical distribution of trees was used for the majority of analyses. However, since the effect of using an empirical tree distribution in discrete trait mapping analyses has not been reported in the literature, I also compared the effect of using empirical tree distributions or joint inference of phylogenies and discrete traits on the inferred number and rate of transitions.

Empirical tree distributions were generated in BEAST under the SRD06 nucleotide substitution model, with a relaxed demographic prior (Bayesian skyride with time-aware smoothing). The MCMC was run over at least 100 million generations for each analysis, with sampling of trees and parameters every 10,000 generations. At least 2 independent runs were conducted for each dataset and set of parameter choices. Examination of the BEAST output files with Tracer indicated that a 10% burnin period was sufficient for convergence to have been achieved in all runs. A subset of 1,000 post-burnin phylogeny samples was obtained for use as empirical tree samples, due to memory constraints. To infer discrete ancestral states along the empirical tree samples, the MCMC was run over 50 million generations, with sampling every 5,000 generations. When joint inference of phylogenies and discrete trait mapping was performed, the MCMC was conducted over 50 million generations, with sampling every 5,000 generations, for each independent run.

5.4.4 Processing and interpretation of discrete trait mapping output

Particular analyses were performed in this chapter to check that the discrete trait mapping methods behaved as expected, and to compare the output of the implementations of Lemey *et al*. (2009) with the Markov jumps methods of Minin and others (Minin and Suchard 2008b; Minin and Suchard 2008a; O'Brien *et al.* 2009; Talbi *et al.* 2009). Such comparisons have not been widely reported in the literature. Custom R scripts were written to post-process the output of the discrete trait mapping analyses, for example to combine the relative rate and indicator information from runs in which BSSVS was implemented. The Cytoscape software (Shannon *et al.* 2003) was used to visualise networks depicting the transition models for the dissemination of character state changes across the phylogenies. Inferred ancestral viral NA subtypes, avian host types and viral pathogenicity were visualised along the branches of BEAST maximum clade credibility (MCC) trees for Eurasian avian H7 HA using FigTree.

5.5 Results

Unless explicitly stated otherwise, the results reported in this section are for discrete trait mapping analyses where empirical tree samples were used, and an asymmetric continuous time Markov chain was employed to model discrete trait transition. BSSVS should be assumed not to have been used unless otherwise stated. Comparisons of the output from different types of discrete trait mapping analysis are described in the text.

5.5.1 Visualising ancestral states upon summary phylogenies

Eurasian avian H7 HA maximum clade credibility trees, onto which discrete ancestral traits had been mapped, were constructed with a posterior probability limit of zero. Branches of the MCC tree were then coloured according to the viral NA subtype (Figure 5.1a), avian host (Figure 5.2a) and viral pathogenicity (Figure 5.3a) at the parental node. Note that discrete trait transitions may have taken place at any point along the branches, but it is not possible to visualise the precise location of changes using currently available software such as FigTree. Although transitions may be overlooked when there are multiple changes along a single branch, colouring the branches according to the node states still captures the discrete trait dissemination across the tree. Similarity in the MCC colourings was observed across discrete trait mapping runs using empirical and non-empirical tree distributions, as well as using symmetric and asymmetric transition models. The ancestral NA subtype at the root of the MCC tree was inferred to be N1 (posterior probability = 0.715) and an LP root was inferred (posterior probability = 0.995). Although the avian host state at the root was inferred to be domestic birds (posterior probability = 0.953), this may be a sampling effect of predominance of earlier sequences from domestic birds and a lack of sampling and sequencing from wild birds until relatively recently. Indeed, the inferred root state was the same as the most basal clade in a phylogeographic analysis of H5N1 viruses (see Figure 1 of Lemey *et al*. 2009). The branch colourings of individual MCMC trees were compared with the inferred number of Markov jumps across the tree and a good concordance was observed.

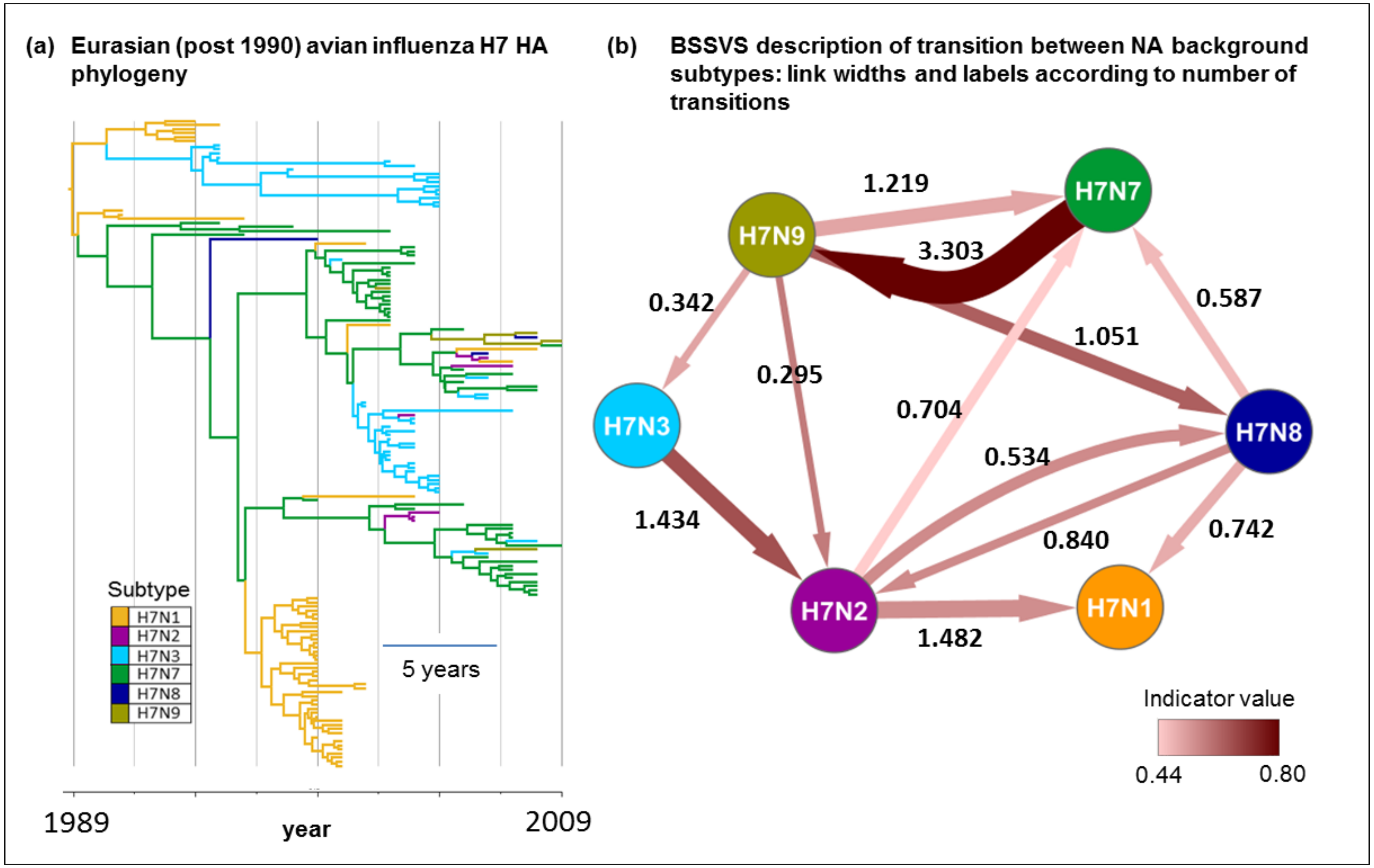

(a) Eurasian (post 1990) avian influenza H7 HA phylogeny
Subtype
H7N1
H7N2
H7N3
H7N7
H7N8
H7N9
5 years
1989
year
2009
(b) BSSVS description of transition between NA background subtypes: link widths and labels according to number of transitions
H7N9
H7N7
H7N3
H7N8
H7N2
H7N1
1.219
3.303
0.342
0.587
1.051
0.295
0.704
0.534
1.434
0.840
0.742
1.482
Indicator value
0.44
0.80

Figure 5.1
Discrete trait mapping of background viral NA subtypes on post-1990 avian Eurasian H7 HA phylogenies. (a) Branches of the BEAST MCC tree were coloured according to inferred ancestral subtypes. (b) Links in the network are significantly non-zero (Bayes factor > 3) rates from the BSSVS analysis, and are coloured according to the indicator value (proportion of time the rate was switched on in the MCMC). Links are labelled by the mean number of transitions from one subtype to another in a Markov jumps analysis, and link widths are proportional to the number of transitions.

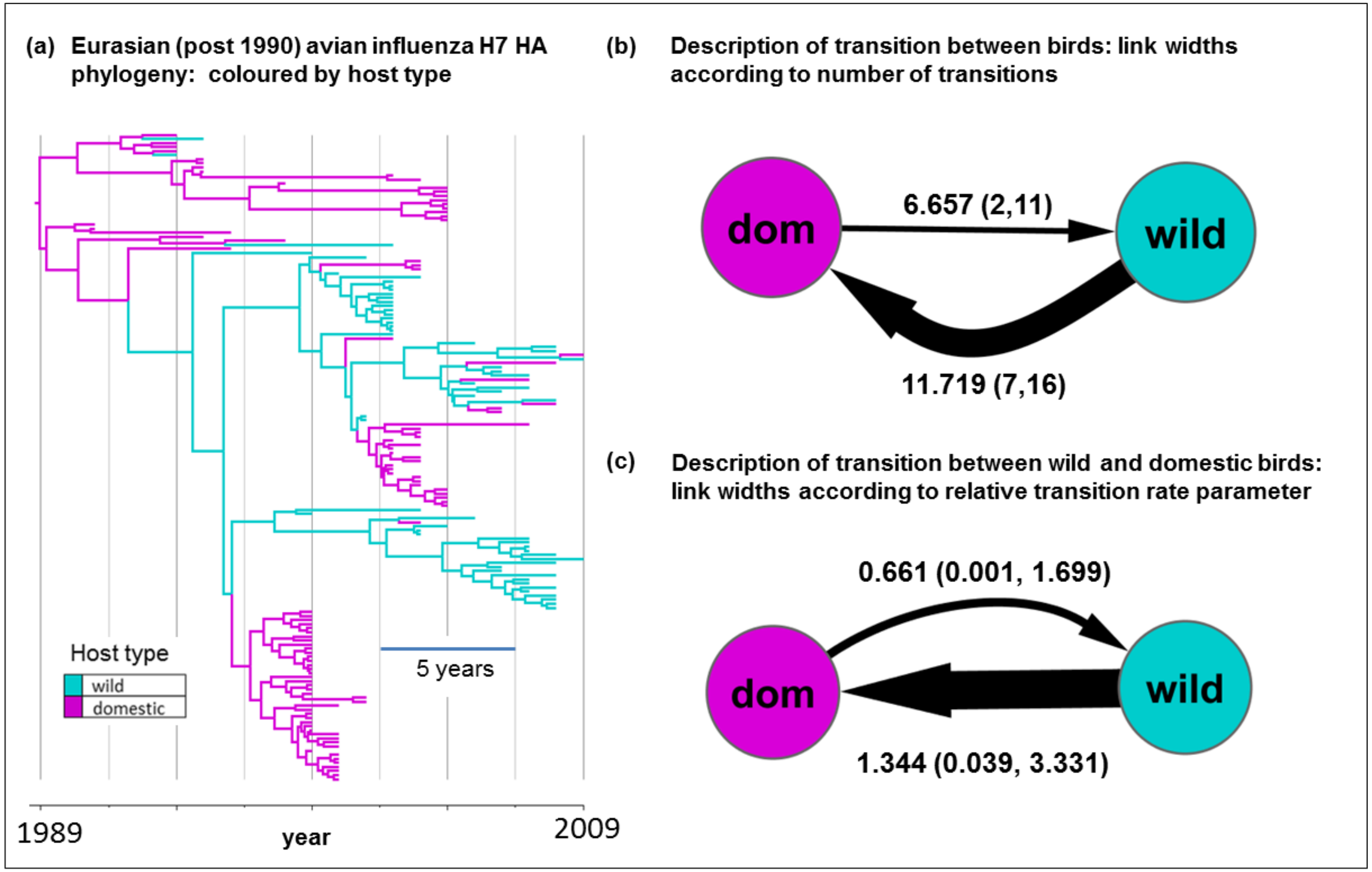

(a) Eurasian (post 1990) avian influenza H7 HA phylogeny: coloured by host type
Host type
wild
domestic
5 years
1989
year
2009
(b) Description of transition between birds: link widths according to number of transitions
dom
wild
6.657 (2,11)
11.719 (7,16)
(c) Description of transition between wild and domestic birds: link widths according to relative transition rate parameter
0.661 (0.001, 1.699)
dom
wild
1.344 (0.039, 3.331)

Figure 5.2: Discrete trait mapping of avian host types on post-1990 avian Eurasian H7 HA phylogenies. (a) Branches of the BEAST MCC tree were coloured according to inferred ancestral hosts (wild or domestic birds). (b) Links are labelled by the mean and 95% HPD limits for the number of transitions from one host type to another, and link widths are proportional to the number of transitions. (c) Links are labelled by the mean (and 95% HPD limits) of the relative instantaneous transition rate from one host type to another, and link widths are proportional to the relative rates. Both rates were switched on in all chain steps in the BSSVS analysis (indicator values of 1), hence the output was the same for BSSVS and non-BSSVS runs.

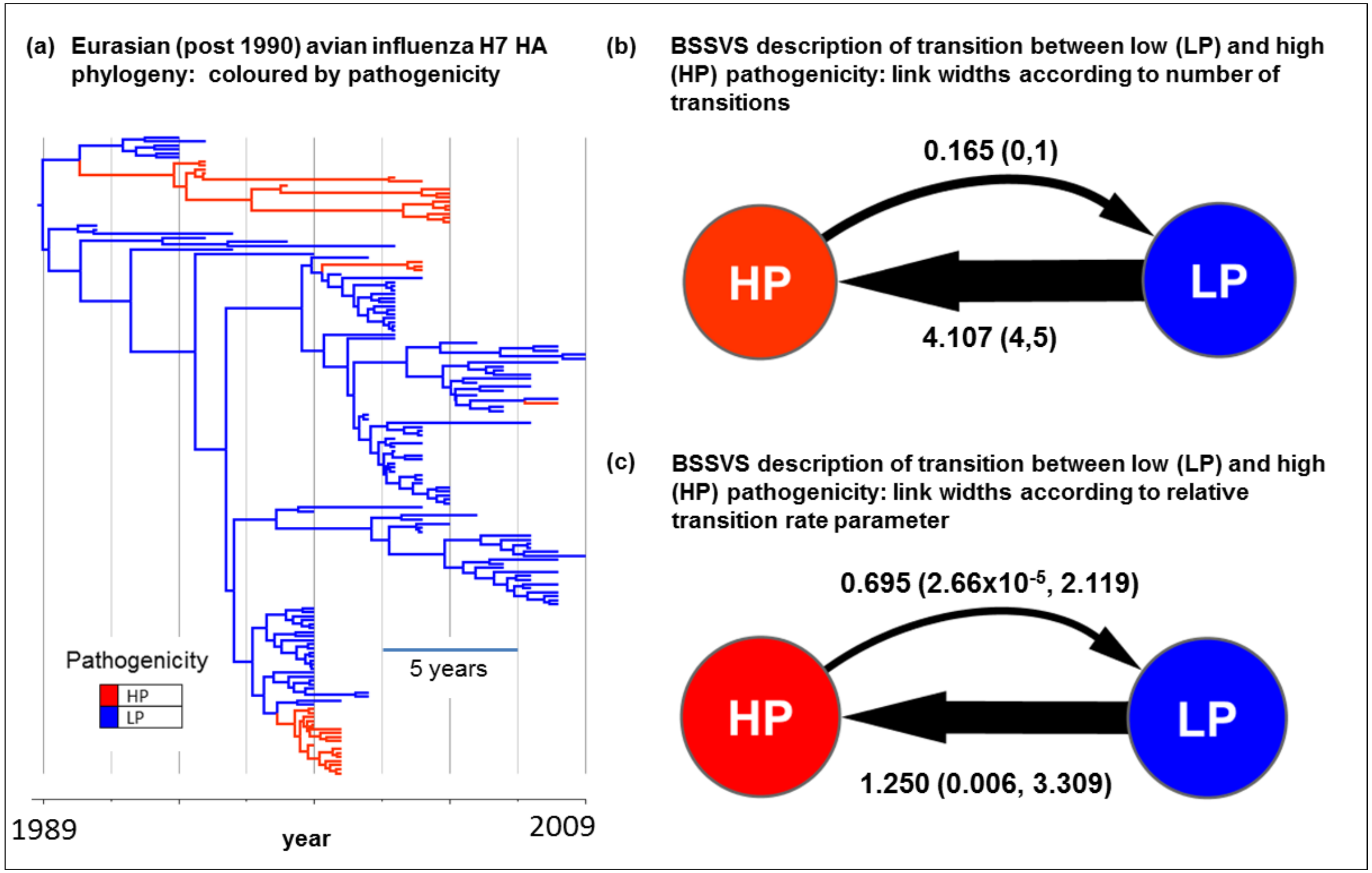

(a) Eurasian (post 1990) avian influenza H7 HA phylogeny: coloured by pathogenicity
Pathogenicity
HP
LP
5 years
1989
year
2009
(b) BSSVS description of transition between low (LP) and high (HP) pathogenicity: link widths according to number of transitions
0.165 (0,1)
HP
LP
4.107 (4,5)
(c) BSSVS description of transition between low (LP) and high (HP) pathogenicity: link widths according to relative transition rate parameter
0.695 (2.66x10^{-5}, 2.119)
HP
LP
1.250 (0.006, 3.309)

Figure 5.3: Discrete trait mapping of viral pathogenicity on post-1990 avian Eurasian H7 HA phylogenies. (a) Branches of the BEAST MCC tree were coloured according to inferred pathogenicity (LP = low pathogenicity; HP = highly pathogenic). (b) Links are labelled by the mean and 95% HPD limits for the number of transitions in pathogenicity, and link widths are proportional to the number of transitions. (c) Links are labelled by the mean (and 95% HPD limits) of the relative instantaneous transition rates between LP and HP, and link widths are proportional to the relative rates. Both rates were switched on in all chain steps in the BSSVS analysis (indicator values of 1), hence the output was the same for BSSVS and non-BSSVS runs.

5.5.2 Instantaneous rate and number of discrete trait transitions

As well as overall rates of discrete trait transition (see Section 5.5.3 and Table 5.3), parameters describing the relative transition rates between particular pairs of states were calculated across the phylogeny samples (Figure 5.4). In addition, the number of transitions (Markov jumps) was calculated between all pairs of states (Table 5.2). An average of 35.7 (95% HPD = 25, 49) transitions between viral NA subtypes were observed across the avian H7 HA phylogenies in the Markov jumps analysis. Six distinct NA subtype states (N1, N2, N3, N7, N8 and N9) were observed at the tips of the phylogenies. Therefore, a total of 5 transitions between NA subtypes would have been observed, even if the H7 HA sequences clustered perfectly according to NA subtype and there had been no reassortment between NA subtypes on the H7 HA background. Any ‘extra’ NA subtype transitions required to map the NA subtypes onto the phylogeny reflect the way in which reassortment has shaped the distribution of NA subtypes at the tips. Similarly, the number and nature of ‘excess’ discrete trait transitions in the host and pathogenicity analyses (where there are two traits in both, so a minimum of one transition would be required) reflects host-switching between wild and domestic birds, and between viruses of high and low pathogenicity, respectively.

Although an overall clock rate for discrete trait transition may be estimated directly in BEAST, this does not allow for comparison of rates between traits where there are different numbers of states, because the minimum number of state changes (see above) is not accounted for. The total number of discrete trait transitions for each trait, minus the minimum number of expected trait changes, was calculated; the number of excess transitions could then be compared between the NA subtype, avian host and pathogenicity analyses, as the discrete traits were being mapped onto the same phylogeny samples in each case. The mean number of excess transitions was much higher for background NA subtype (30.65, 95% HPD = 20, 44) than for avian host (17.39, 95% HPD = 12, 24) or viral pathogenicity (3.28, 95% HPD = 3, 5). This indicates that reassortment of different NA subtypes onto the H7 HA background has occurred substantially more frequently than switches in avian host or viral pathogenicity over the evolutionary history of Eurasian avian H7 influenza, based upon this sample of sequences from post-1990. 95% HPD intervals for the number of excess trait transitions across the phylogeny samples were significantly non-zero for the viral NA subtype, host and pathogenicity-mapping analyses, confirming the presence of reassortment with NA, host-switching between wild and domestic birds and repeated emergence of viruses of different pathogenicity amongst avian H7 HA.

Some similarity was observed in the pattern of relative instantaneous transition rate parameters between pairs of viral NA subtypes for the non-BSSVS analysis and the BSSVS runs analysed in two ways (including or excluding rates which were ‘switched off’ in the MCMC) (Figure 5.4). In the non-BSSVS and BSSVS analyses, the lower limit of the 95% HPD intervals lay towards zero for all rates except the rate of instantaneous transition from N7 to N3 in the non-BSSVS analysis (Figure 5.4a). However, differences in the upper 95% HPD limits were observed between the non-BSSVS and BSSVS analyses. Using BSSVS appeared to have the effect of pushing the rates with the lowest means and upper 95% HPD limits towards zero, thus allowing a higher degree of differentiation between the instantaneous pairwise transition rates (Figure 5.4b and Figure 5.4c).

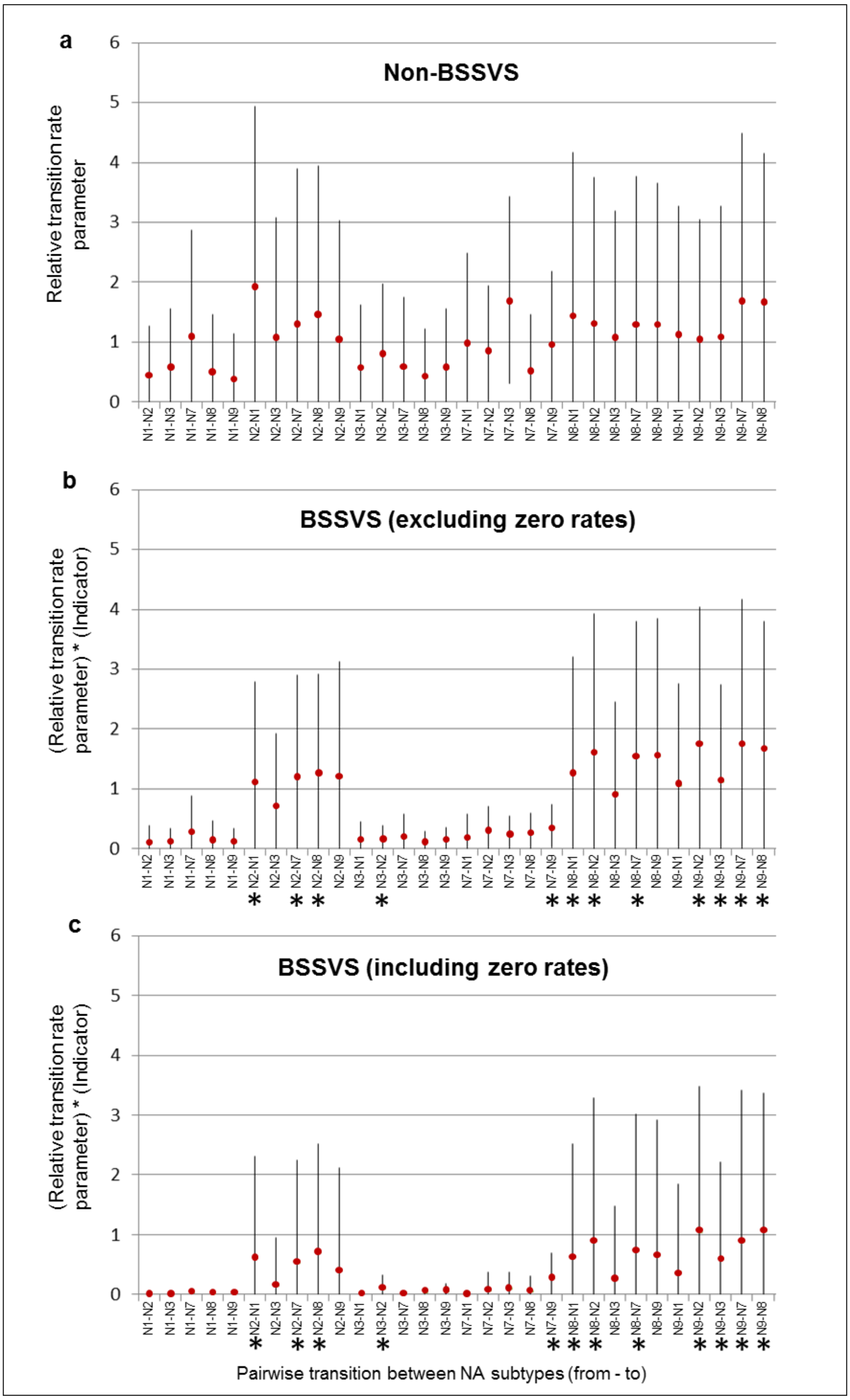


Figure 5.4

Relative instantaneous transition rate parameters between background NA subtypes. Plot (a) shows transition rates from a non-BSSVS analysis. In (b), only the non-zero rates from the BSSVS analysis are analysed, whereas in (c) the distributions include zero and non-zero rates, obtained by multiplying the relative rate parameters by the indicator values. 95% HPD intervals are shown as vertical lines, with red dots indicating the mean. BSSVS rates which were significant under Bayes factor testing are labelled with asterisks.

Under the parameterisation described by Lemey *et al*. (2009), the relative transition rate parameters have a mean of one in the non-BSSVS analysis. Although the raw relative rate parameters in a BSSVS analysis also have a mean of one, when they are post-processed by multiplying the relative rate parameters by the rate indicators (to account for rates being switched off at points along the MCMC chain) this is no longer the case. As expected, the post-processed relative rate parameters were lower when averaged across the whole of the MCMC chain (including rates which were set to zero as they were switched off in the MCMC) compared to averaging only over non-zero rates. When the relative rate parameters were multiplied by the corresponding overall clock rates in order to scale the pairwise transition rates to the same units as the phylogeny (i.e., years), the pattern of pairwise transition rates within an analysis (BSSVS or non-BSSVS) was the same as for the relative rate parameters (compare Figure 5.4 and Appendix C, Figure C2). However, large differences in the overall clock rates from the BSSVS and non-BSSVS analyses led to substantially higher means and upper 95% HPD limits for the pairwise instantaneous transition rates in the BSSVS analysis than for the non-BSSVS analysis. The higher mean and upper 95% HPD limit between the subtype transition clock rates for the BSSVS than the for non-BSSVS runs (non-BSSVS: mean = 0.184 transitions/year, 95% HPD interval = [7.00x10^{-2}, 0.334]; BSSVS: mean =1.176 transitions/year, 95% HPD interval = [0.162, 2.562]) is surprising, given that an aim of BSSVS is to reduce noise in the model of character dissemination across the phylogeny. The fact that the overall transition clock rate from the non-BSSVS analysis appeared consistent with the number of transitions across the tree in the Markov jumps analysis (see Section 5.5.3 below) suggests that it is the BSSVS output which might require additional processing or consideration in the future.

In the BSSVS analysis, statistical support (Bayes factor > 3, following Lemey *et al*. (2009)) was found for including 12 directed pairwise transition rates in the model of NA subtype dissemination across the phylogeny samples. All of the significant rates from the BSSVS analysis involved the three most rarely occurring NA subtypes (N2, N8 and N9) which between them only accounted for approximately 8% of the H7

HA sequence data. The transition rate from NA subtype N7 to N3, which each accounted for almost one third of the sequences in the dataset, was not found to be significantly non-zero in the BSSVS analysis. However, in the non-BSSVS analysis the relative instantaneous transition rate parameter from N7 to N3 was one of the highest, and was the only relative rate parameter with a lower 95% HPD limit which was substantially above zero (Figure 5.4a). Furthermore, the highest number of transitions (6.513, 95% HPD interval 3,9) from the Markov jumps analysis was from subtype N7 to N3 (Table 5.2). 10 out of the 12 rates which were significant under BSSVS had high means and upper 95% HPD limits for their relative rate parameters, whereas only two significant rates (N3 to N2 and N7 to N9) had low means and upper 95% HPD limits (Figure 5.4). It is also notable that all 10 significant rates with high means and upper 95% HPD limits involve transitions from subtypes which occurred rarely at the tips of the taxa (N2, N8 and N9; 6, 3 and 4 sequences respectively, out of 159), whilst only two significantly non-zero rates under BSSVS were transition rates from the more frequently occurring subtypes (N3 and N7; 44 and 46 sequences respectively), and these had the lower relative instantaneous transition rate means and upper 95% HPD intervals under BSSVS. These findings suggest that the rates which are found to be significantly non-zero under BSSVS could be highly influenced by the relative frequencies of trait states at the tips of the tree, and therefore also by sampling.

For the avian host analysis, both the relative instantaneous transition rate and the overall number of transitions were higher from wild to domestic birds than from domestic birds to wild birds (Figure 5.2b-c; Table 5.2). This is not surprising, given the role of wild birds as a reservoir population for avian influenza viruses (see Chapter 1, Section 1.4). However, some transitions from domestic birds to wild birds were required to explain the distribution of sequences from different avian hosts at the tips of the phylogeny. Some overlap in the HPD intervals for transition from wild to domestic hosts, and from domestic to wild hosts, was observed for both the relative rate parameters and number of Markov jumps. Because the transition rates from wild to domestic birds and from domestic to wild birds were switched on in every MCMC state in the BSSVS analyses, the overall transition clock rates and relative instantaneous transition rate parameters were the same for the asymmetric BSSVS and non-BSSVS analyses, and both transition rates were found to be significantly non-zero under BSSVS.

The relative instantaneous transition rate from LP to HP viruses was higher than from HP to LP viruses (Figure 5.3c), although there was substantial overlap between the 95% HPD intervals for the two rates. In the Markov jumps analysis, essentially all of the transitions were from LP to HP (4.107, 95% HPD interval 0,1) rather than from HP to LP (0.165, 95% HPD interval 0,1) (Figure 5.3b). This is in line with the notion that highly pathogenic avian influenza viruses arise sporadically from low pathogenic forms, and the fact that a literature search yielded no documented evidence of highly pathogenic avian viruses becoming low pathogenic *in vivo*.

	Mean number of jumps	Lower 95% HPD limit	Upper 95% HPD limit	Significance under BSSVS
wild - domestic	11.719	7	16	*
domestic - wild	6.657	2	11	*
LP - HP	4.107	4	5	*
HP - LP	0.165	0	1	*
H7N1 - H7N2	0.429	0	2	
H7N1 - H7N3	0.862	0	3	
H7N1 - H7N7	2.725	0	8	
H7N1 - H7N8	0.659	0	2	
H7N1 - H7N9	0.288	0	2	
H7N2 - H7N1	1.482	0	4	*
H7N2 - H7N3	0.333	0	2	
H7N2 - H7N7	0.704	0	3	*
H7N2 - H7N8	0.840	0	3	*
H7N2 - H7N9	0.306	0	2	
H7N3 - H7N1	0.749	0	3	
H7N3 - H7N2	1.434	0	3	*
H7N3 - H7N7	0.814	0	4	
H7N3 - H7N8	0.285	0	1	
H7N3 - H7N9	0.699	0	2	
H7N7 - H7N1	3.325	0	7	
H7N7 - H7N2	2.806	0	5	
H7N7 - H7N3	6.513	3	9	
H7N7 - H7N8	1.280	0	4	
H7N7 - H7N9	3.303	1	6	*
H7N8 - H7N1	0.742	0	3	*
H7N8 - H7N2	0.534	0	2	*
H7N8 - H7N3	0.236	0	1	
H7N8 - H7N7	0.587	0	3	*
H7N8 - H7N9	0.504	0	2	
H7N9 - H7N1	0.446	0	2	
H7N9 - H7N2	0.295	0	2	*
H7N9 - H7N3	0.342	0	2	*
H7N9 - H7N7	1.219	0	4	*
H7N9 - H7N8	1.051	0	3	*

Table 5.2
Number of discrete trait transitions on Eurasian avian H7 HA phylogenies. The mean and 95% highest posterior density (HPD) intervals of the number of transitions between different states (wild and domestic avian hosts, viruses of high and low pathogenicity and pairs of NA subtype backgrounds) are reported. An asymmetric transition model was implemented so that direction could be inferred between pairs of states: for example, 'wild-dom' refers to a transition along the tree from wild to domestic avian host, and 'dom-wild' refers to a transition from domestic to wild. Rows marked with an asterisk (*) denote transition rates which were found to be significantly non-zero under a BSSVS analysis.

5.5.3 Technical investigation of discrete trait mapping methods in BEAST

Concordance between the output from different BEAST discrete trait mapping analyses and models (Table 5.3) was investigated. When asymmetric CTMC transition models were employed, the mean instantaneous substitution rates for each trait (NA subtype, host type and pathogenicity) were consistent across runs using empirical and non-empirical trees. This suggested that there was no discernible difference in the inferred overall transition rate when phylogeny inference and discrete trait mapping were performed separately or jointly. Individual BSSVS relative rate parameters between pairs of states were also consistent between empirical and non-empirical runs (not shown). Although the mean transition rates appeared to be slightly different in the symmetric runs compared to the asymmetric runs, the substantial overlap between the HPD intervals indicated that such differences were not significant. There also was a high level of agreement in the mean number of transitions (Markov jumps) between analyses using symmetric or asymmetric trait-transition models, as well as between using an empirical distribution of trees and concomitant inference of phylogenies and discrete ancestral traits. The lower limits of the 95% HPD intervals for the number of Markov jumps across the tree samples was more consistent between analyses than the upper 95% HPD limit. Overall, these results indicate that the total number of Markov jumps is robust to different choices of model and analysis.

Trait	Run	Mean no. Markov jumps	Lower 95% HPD limit (jumps)	Upper 95% HPD limit (jumps)	Mean transition clock rate	Lower 95% HPD limit (rate)	Upper 95% HPD limit (rate)
NA subtype	asymmetric empirical	35.7	25	49	0.183	7.15E-02	0.3257
NA subtype	asymmetric non-empirical	34.4	25	47	0.184	7.00E-02	0.3336
NA subtype	symmetric empirical	33.822	25	43	0.138	6.63E-02	0.2248
Host	asymmetric empirical	18.32	13	25	7.63E-02	3.58E-02	0.1217
Host	asymmetric non-empirical	17.77	13	23	7.67E-02	3.61E-02	0.1214
Host	symmetric empirical	19.109	14	27	7.72E-02	3.77E-02	0.1245
Pathogenicity	asymmetric empirical	4.28	4	6	1.83E-02	2.99E-03	2.92E-02
Pathogenicity	asymmetric non-empirical	4.26	4	6	1.83E-02	2.83E-03	3.88E-02
Pathogenicity	symmetric empirical	4.39	4	6	2.06E-02	5.05E-03	3.97E-02

Table 5.3
Number and rate of discrete trait transitions across Eurasian avian H7 HA phylogenies (non-BSSVS). Discrete trait mapping of viral NA subtype (N1, N2, N3, N7, N8 or N9), avian host (wild or domestic) and viral pathogenicity (LP or HP) was performed upon Eurasian avian H7 HA phylogenies constructed from viral sequences sampled after 1990. Bayesian stochastic search variable selection (BSSVS) was not implemented in these runs. The mean instantaneous transition rate from the discrete trait mapping analysis was reported for the different analyses, employing symmetric and asymmetric continuous time Markov chain transition models and either using an empirical sample of trees or performing a joint inference of phylogenies and discrete ancestral traits. The mean and upper and lower limits of the 95% HPD interval are reported. After discrete trait mapping, a Markov jumps analysis was performed upon BEAST phylogeny samples to infer the number of NA subtype transitions across the trees. The number of transitions reported across the tree was robust to the use of symmetric or asymmetric models, and to the joint inference of phylogenies and ancestral discrete traits or the use of an empirical set of phylogeny samples.

The expected total number of transitions across the phylogeny under the inferred overall instantaneous transition rate (calculated from the overall subtype clock rate multiplied by the tree length) was compared to the observed number of changes from the Markov jumps counting. For the NA subtype, avian host and pathogenicity mapping analyses, an overall concordance was observed between the expected and observed numbers of substitutions (Appendix C, Table C3; Appendix C, Figure C3). For example, for the Eurasian post-1990 dataset under a symmetric transition model, the mean expected number of subtype transitions was 36.36 (95% HPD: 16.81, 58.96) and the mean observed number of subtype transitions was 38.82 (95% HPD limits = 25, 43). The most departure between the observed and expected number of transitions was for the NA subtype mapping analysis with an asymmetric transition model, which may be visualised as departure from the line *y*=*x* on the scatterplot of observed against expected numbers of jumps (Appendix C, Figure C3b). It is possible that this has arisen from calculating the expected number of transitions using only the overall transition clock rate and tree length, rather than adding together 'component-wise' the expected numbers of jumps between pairs of states and accounting for differences in the time spent along the branches in each state. Nonetheless, the method for obtaining estimated number of transitions across the tree used here is a useful indicator that the overall transition clock rate is meaningful in terms of the number of transitions observed across the tree. The HPD intervals were narrower for the Markov jumps counts than for the expected number of transitions calculated from the transition rate and tree length; it is likely that this arises from the way in which the Markov jumps analysis firstly infers the infinitesimal transition rates in a conventional manner, but also incorporates the empirical state frequencies from the dataset (O'Brien *et al*. (2009); Philippe Lemey, personal communication).

The output from BSSVS, non-BSSVS and Markov jumps analyses was compared for discrete trait mapping of viral NA subtypes on Eurasian avian H7 HA phylogenies. No obvious relationship was observed between the number of transitions from one state to another and the corresponding instantaneous relative rate parameter for transition between these states (Appendix C, Figure C4a-c; Appendix C, Table C4).

This may be because the time spent along the tree in the state being moved from (equivalent to the ‘dwell times’ in Chapter 4) is not accounted for; with additional processing, such information may be obtained in the future from a ‘Markov rewards’ analysis (Minin and Suchard 2008b). In general, pairwise transitions with a higher relative rate parameter were more likely to be switched on in the MCMC (and thus be significantly non-zero in the BSSVS analysis) (Appendix C, Figure C4d-f; Appendix C, Table C4), although it may be observed that some transitions with a low relative rate parameter in fact have a high indicator value in the BSSVS analysis (Appendix C, Figure C4e-f). Finally, relative rate parameters calculated from a BSSVS analysis were found to be positively correlated with the relative rate parameters from a non-BSSVS analysis, both over the whole MCMC chain and when only non-zero rates were considered (Appendix C, Figure C4g-h). A strong positive relationship was observed between BSSVS relative rate parameters averaged over all states of the MCMC (with rates at which the indicator was 0 being set to zero) and the BSSVS relative rate parameters averaged over only the non-zero states (Appendix C, Figure C4i).

5.6 Discussion

In this chapter, discrete trait mapping methods which have traditionally been used for phylogeographic analysis were employed to investigate discrete trait transitions across H7 avian influenza HA phylogenies. Starting from information at the tips of the phylogenies, viral NA subtypes were mapped onto the tree samples in BEAST to quantify reassortment between H7 HA and NA segments of different subtypes. This is the first time that such an approach has been applied to avian influenza data. Transitions between wild and domestic hosts, and between viruses of different pathogenicities, were also considered. Previous researchers performing discrete trait mapping in BEAST have compared relative instantaneous transition rate parameters between pairs of discrete states (e.g. Lycett *et al*. (2012)), whilst others have calculated numbers of transitions between pairs of states (e.g. Talbi *et al*. (2009));

however, studies comparing the output of such analyses have not been performed. The results of this chapter imply that the total number of discrete trait transitions from a Markov jumps analysis is a robust measure of discrete trait transition across phylogenies, whereas the behaviour observed in this study indicates that the BSSVS output may require further consideration. In contemplating the relative merits of different discrete trait mapping methods for quantifying reassortment from phylogenies, this work preceded the study described in Chapter 6. Following this chapter, and given known difficulties in informing relative transition rates (Philippe Lemey, personal communication), a Markov jumps approach is taken in Chapter 6 for investigating inter-subtype recombination in HIV-1 group M using the total number of subtype transitions across phylogenies from different ends of the genome.

The discrete trait mapping analyses reported in this chapter revealed reassortment between H7 influenza HA and different NA subtypes to occur substantially more frequently than host-switching between wild and domestic birds, which in turn occurred more often than transitions between LP and HP viruses. On average, 30.65 (95% HPD interval = 20,44) NA subtype transitions occurred across the tree per year as a result of reassortment, compared to 17.39 (95% HPD interval = 12,24) host switches per year and 3.28 (95% HPD interval = 3,5) changes in viral pathogenicity, indicating reassortment to be a major evolutionary factor in generating avian influenza virus diversity. This is the first time that reassortment between HA and a large number of NA subtypes has been quantified across avian influenza phylogenies. Although there was not a large amount of variation in transition rates from one NA subtype to another in the non-BSSVS analysis, implementing BSSVS made it easier to distinguish between rates. A notable observation from the BSSVS analysis for NA subtype mapping was that the transition from subtype N7 to N3, which was a relatively high, non-zero rate in the non-BSSVS analysis, and which accounted for the highest number of transitions in the Markov jumps analysis, was not found to be significantly non-zero under BSSVS. Such findings indicate that some care may be required when interpreting the results of BSSVS analyses.

It is possible that the ‘significant’ rates of diffusion under BSSVS could be highly influenced by rare events or sampling effects, and that they should not be over-interpreted biologically. For example, if there is only one sequence of trait state A in a dataset and that taxon clusters with the sequences of state B with a high posterior probability, the transition rate between states A and B will necessarily be significant in order to explain the existence of state A at the tip of the tree. In this study, all of the rates inferred to be significantly non-zero under BSSVS involved at least one of the rarest three background NA subtype states, and the most highly-connected nodes were those corresponding to the rarest three states (Figure 5.1). In future, information about the number of taxa corresponding to each state (such as the number of sequences for the start and end states, or the product of the number of sequences corresponding to the start and end states) could be incorporated into a generalised linear model as predictor variables (Philippe Lemey and Nuno Faria, personal communication), as a method for investigating whether the discrete trait dispersal model is being unduly influenced by the number of sampled sequences in each state. It must also be noted that the discrete trait transition models assume the state frequencies to be homogenous over the tree (i.e. over time). This assumption may be violated in situations such as when the sampling strategy has changed over time. In future, models which allow different transition rate matrices for different time-periods may be developed.

A faster relative instantaneous transition rate was observed from wild birds to domestic birds than from domestic birds to wild birds. Although this has been hypothesised before from avian influenza phylogenies, by Lebarbenchon and Stallknecht (2011), their analysis only involved colouring the tips of the phylogeny by host type and counting the number of times that wild and domestic sequences formed sister lineages. In contrast, the quantitative analysis performed in this chapter allows the rate of host-swapping to be compared to the rate of dissemination of other characters across the phylogenies, and also for directionality and the relative rate or number of transitions from one host type to another to be inferred. In the future, additional techniques for inferring patterns of pathogen transmission from mapping discrete traits upon phylogenies could allow formal hypothesis testing to be performed. For example, a null distribution for the expected number of transmission events between different avian host types could be generated by performing a large number of randomisations of the states at the tips of the phylogeny and then counting the number of each type of transition required to explain the distribution of states at the tips for each randomisation. The observed number of transitions in the phylogenies with the non-randomised tips could then be compared to the null distribution, to test for departures from random transmission and identify higher- or lower-than-expected mixing (Goldberg 2003). In principle, it is possible to perform a large number of randomisations of the tip states and produce an individual BEAST xml file for each randomisation, then to perform the Markov jumps transition counting for each xml file using an empirical distribution of trees. However, there is not currently an in-built method for randomising tips and producing such a null distribution in BEAST.

The finding of a non-zero transition rate from domestic to wild avian hosts in the BSSVS analysis could be a result of transmission of influenza viruses from poultry back into wild avian populations, as has been witnessed for H5N1 in Asia (Chen *et al.* 2005; Feare 2010). However, it may also be an artefact of sampling, in particular under-sampling of influenza from wild birds. Future investigations with well-sampled sequence data from both wild and domestic hosts could consider the relative contribution of transmission from wild to domestic birds and spill-over back from domestic to wild birds. In addition, joint discrete trait mapping of host and NA subtype could be performed to test for association between rates of reassortment and avian influenza in wild or domestic birds, for example to examine whether reassortment occurs more frequently in wild avian populations. Furthermore, the rate of reassortment between different subtypes is likely to be related to the opportunity for reassortment between different subtypes. Rates of transition across the phylogenies between particular pairs of subtypes could thus be reconciled with subtype prevalence in the population, if such data were available.

Future applications of discrete trait mapping on influenza phylogenies could be used to formally test whether different HA and NA subtype combinations arise randomly in the wild avian reservoir as a result of reassortment, or whether there is statistical evidence that certain HA/NA subtype combinations are preferred. For example, separate discrete trait mapping analyses of HA subtypes and NA subtypes could be performed upon a set of BEAST phylogeny samples constructed from internal influenza gene segments (e.g. PB2 sequences from North American wild birds, see Figure 5.5). This would lead to individual transition matrices for transition between different HA subtypes and between different NA subtypes which could then be used to simulate HA and NA subtypes along the phylogenies, starting from the inferred state at the root of the tree, under the assumption that the HA and NA subtypes have evolved independently. The number of each HA/NA subtype combination at the tips of the tree may then be counted. By performing such simulations many times, and across a distribution of phylogeny samples, a 'null distribution' for the number of each HA/NA subtype combination under independent reassortment could be obtained, against which the actual number of each HA/NA subtype combination in the avian influenza dataset could be compared. The effect of sampling strategy, sample coverage and the method used to down-sample available sequence data (for computational reasons or to avoid bias from over-sampling) has not been formally addressed in majority of published phylogeographic or discrete trait analyses using BEAST, but should also be considered in the future.

Although this chapter focused on reassortment between H7 HA and different NA subtypes, discrete trait mapping could also be used to assess the extent of reassortment between all influenza RNA segments. Whilst no clear pattern of reassortment between segments has been found by visually comparing phylogenies for avian influenza viruses from wild birds (Dugan *et al.* 2008), visual evidence from phylogenies suggests that particular constellations of polymerase segments PB1, PB2 and PA persist in swine influenza (Vincent *et al.* 2008), as well as in viruses such as H5N1 genotype B which are adapted to domestic poultry (Vijaykrishna *et al.* 2008a). Such hypotheses could be tested quantitatively using discrete trait mapping, by labelling the tips of the tree according to how they cluster in the tree for one segment, then calculating how many discrete state changes are required to map these same tip labels onto the tree for another segment.

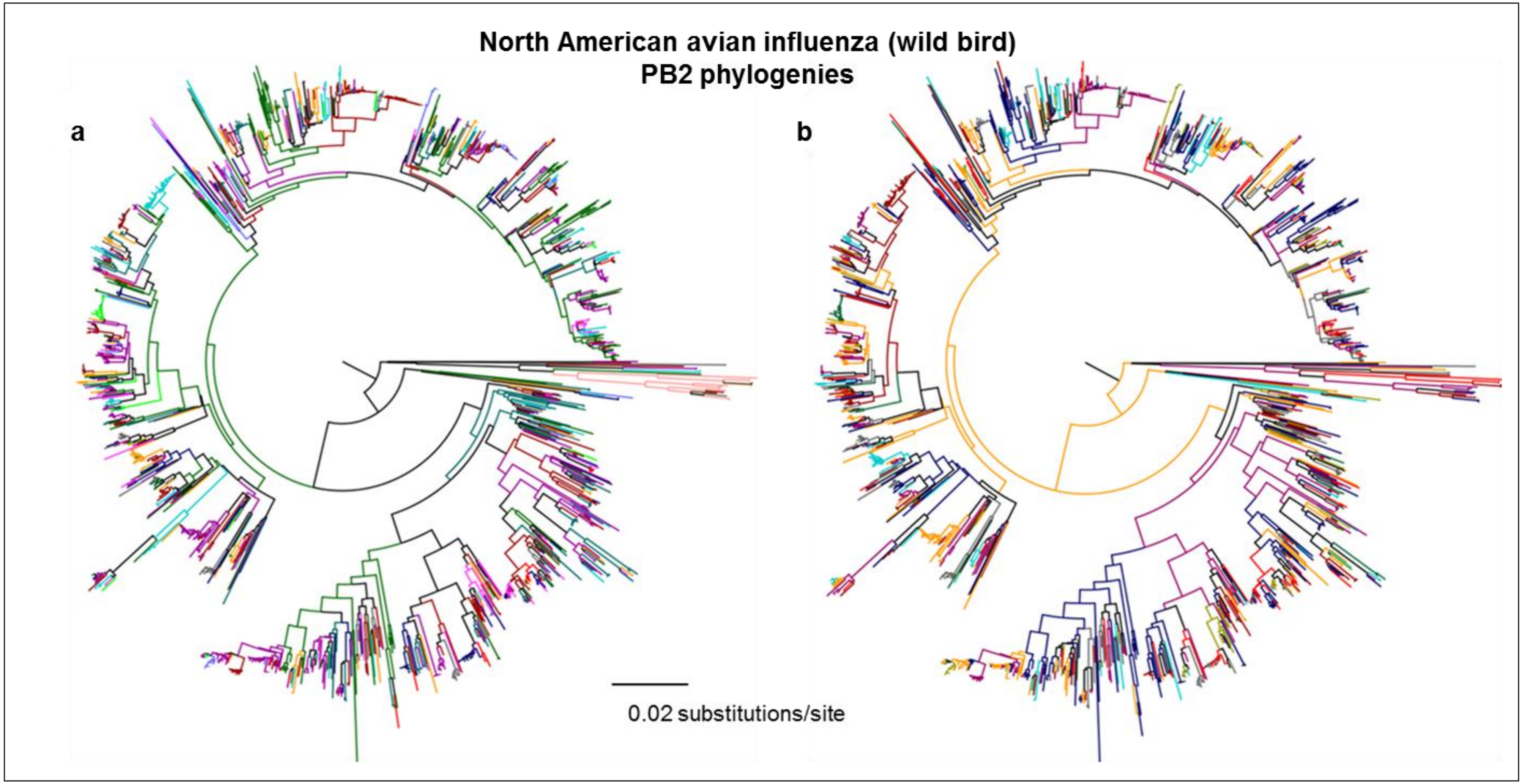


Figure 5.5
Global avian influenza PB2 phylogenies from wild birds, coloured by viral HA or NA subtype. (a) A neighbour-joining phylogeny was constructed from avian influenza PB2 sequences from wild birds. Tips of the tree were labelled by HA subtype and the tree was coloured according to HA subtype using parsimony mapping in FigTree. In (b), the same PB2 phylogeny is shown, but instead it is coloured according to the NA subtype of the virus.

Chapter 6

Quantifying inter-subtype recombination amongst early HIV-1 group M in Kinshasa

6 Quantifying inter-subtype recombination amongst early HIV-1 group M in Kinshasa

6.1 Chapter summary

Western Central Africa has been suggested as the epicentre of the HIV-1 epidemic and contains the full spectrum of HIV-1 group M subtype diversity. Previous analyses of early HIV sequences from Kinshasa in the Democratic Republic of the Congo (DRC) have revealed that sequences from a number of patients fall in different phylogenetic positions in trees constructed using sequences from different ends of the genome. Such patterns have been attributed to recombination between viruses of different subtypes. In this chapter I undertake a detailed phylogenetic analysis of HIV-1 isolates from Kinshasa in 1984 to investigate inter-subtype recombination. This is important for quantifying the ancestral contribution of recombination to HIV diversity, as well as for understanding the extent to which phylogenetic studies which do not account for recombination may be confounded by its presence. I develop a statistic for measuring the amount of inter-subtype recombination detectable from phylogenies from different parts of the genome and consider how it may be rescaled to estimate the rate at which inter-subtype recombination events have arisen across the trees.

6.2 Chapter aims

- Perform phylogenetic analyses to investigate inter-subtype recombination in Kinshasa, using sequences from opposite ends of the HIV-1 genome (encoding the *gag* p17 and *env* gp41 regions)
- Develop a statistic for quantifying inter-subtype recombination by comparing phylogenies for different parts of the genome

- Consider how to rescale the recombination statistic to obtain an estimate of the inter-subtype recombination rate of HIV-1 group M in Kinshasa from across the phylogeny

6.3 Introduction

The human immunodeficiency virus (HIV) belongs to the family *Retroviridae*, within which it is a member of the lentivirus genus (Freed and Martin 2007). HIV is a chronic infection which, in the long term, is associated with a significant decline in CD4$^+$ lymphocytes and progression to the acquired immune deficiency syndrome (AIDS), resulting in death. The AIDS epidemic came to prominence in 1981 and was first associated with a virus by Barré-Sinoussi *et al.* (1983). HIV is closely related to the simian immunodeficiency virus (SIV). HIV infections are caused by one of two immunologically distinct viruses, HIV type 1 (HIV-1) or type 2 (HIV-2). Phylogenetic studies have indicated that independent cross-species transmissions of SIV from chimpanzees in central Africa and sooty mangabeys in Western Africa were responsible for the emergence of HIV-1 and HIV-2 respectively (Gao *et al.* 1992; Gao *et al.* 1999).

HIV-1 isolates separate into four phylogenetically distinct groups: M, N and O, believed to represent three independent transmissions of SIV from chimpanzees to humans (Gao *et al.* 1999; Sharp *et al.* 2001) and group P, which is closely related to strains from gorillas (Plantier *et al.* 2009). HIV-1 group M viruses are the most common and are distributed globally, being responsible for over 95% of HIV infections worldwide (Baird *et al.* 2006). Within HIV-1 group M, a number of subtypes (A, B, C, D, F, G, H, J and K) have been identified, which cluster phylogenetically (Robertson *et al.* 2000) and which have approximately equal genetic distances between them (Rambaut *et al.* 2004). Note that assigning subtypes on a purely phylogenetic basis is different to defining the influenza HA and NA subtypes, which are immunologically distinct.

The core of the HIV-1 virion is shaped by the capsid protein (CA), within which lie two copies of the single positive strand of RNA which encodes the HIV-1 genome. The HIV-1 genome is approximately 9.7 kilobases in length (Ratner *et al.* 1985) and there are three major genes, *gag*, *pol* and *env* (Figure 6.1). The *gag* gene encodes the matrix (MA), capsid and nucleocapsid proteins. The *pol* gene encodes the reverse transcriptase, protease and integrase enzymes required for viral replication. The viral envelope gene, *env*, encodes the single gp160 protein which is cleaved into the gp120 external surface protein (SU) and the gp41 trans-membrane protein (TM), embedded in a lipid membrane originating from an infected host cell. The gp120 and gp41 surface glycoproteins comprise the head and stem, respectively, of the trimeric spikes which protrude at the virion surface and attach to host cells during fusion. The p17 protein is a matrix protein, whose main function is structural.

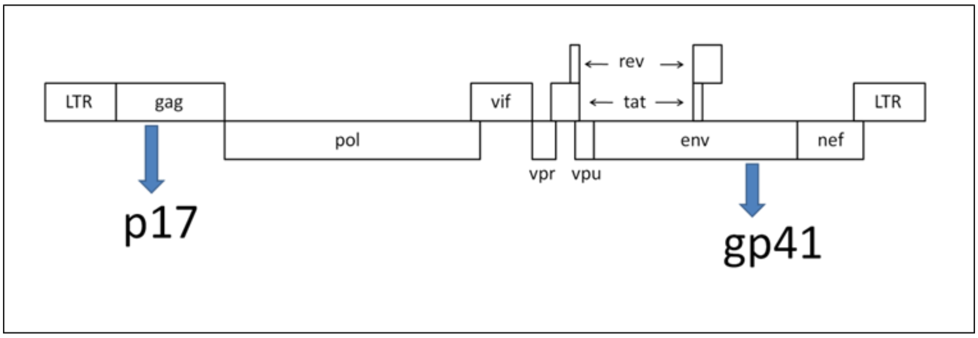


Figure 6.1
Schematic diagram showing the structure of the HIV-1 genome. The major *gag*, *pol* and *env* genes may be observed. The p17 and gp41 proteins (sequences encoding for which are examined in this chapter) are encoded at the 5´ end of *gag* and at the 3´ end of *env* respectively. The HIV-1 genome is approximately 9.7 kilobases long.

HIV-1 is a rapidly-evolving pathogen, with a high mutation rate resulting from an error-prone replication cycle (Preston *et al.* 1988; Roberts *et al.* 1988) and short generation time (Perelson *et al.* 1996). HIV-1 virions contain two copies of the single positive strand of RNA which encodes the HIV genome. Viruses which are detectable as recombinants can arise though template-switching during reverse transcription, when more than one genetically distinct virus is harboured in an infected cell at the same time (Hu and Temin 1990). *In vitro* studies have estimated a minimum of 2.8 crossover events per genome per round of replication: an order of magnitude higher than the rate of mutation, which has been estimated at between 3 x10^{-5} and 8x10^{-5} per nucleotide site (thus approximately 3x10^{-1} and 8x10^{-1} per genome per replication cycle for a virus with a 9.7kb genome) (Zhuang *et al.* 2002). Mutation and recombination can thus both play a significant role in generating HIV diversity (Rambaut *et al.* 2004; Onafuwa-Nuga and Telesnitsky 2009).

Recombination in HIV-1 can occur between viruses of the same subtype (Liu *et al.* 2002; Yang *et al.* 2005b) and between viruses of different subtypes (Robertson *et al.* 1995). Multiple-infection of individuals with viruses of different subtypes, a prerequisite for inter-subtype recombination, may result from a single transmission of genetically different viruses or a subsequent HIV infection acquired by an already infected individual. Inter-subtype recombination was identified as a major mechanism for the generation of HIV-1 group M diversity by Robertson *et al*. (1995), who reported numerous individuals from whom sequences from the *gag* and *env* regions (i.e. from opposite ends of the genome) were of different subtypes based upon phylogenetic analysis. Inter-subtype recombinant viruses with the same breakpoints which are known to have caused infection in three or more epidemiologically unlinked individuals are known as circulating recombinant forms (CRFs) (Robertson *et al.* 2000). At least 50 CRFs have now been characterised (see www.hiv.lanl.gov) and inter-subtype recombinant viruses are thought to account for more than 20% of HIV cases worldwide (Gao *et al.* 2011).

Western Central Africa contains the full spectrum of HIV subtype diversity (Vidal *et al.* 2000). The oldest known HIV-1 sequence (the partial env *sequence* ZR59), identified by retrospective testing of stored samples, dates from 1959 in the Democratic Republic of the Congo (DRC) (Zhu *et al.* 1998). Phylogenetic analyses which include ZR59 and a sequence from the DRC in 1960 (DRC60) have placed the most recent common ancestor of HIV-1 group M towards the start of the twentieth century (Worobey *et al.* 2008). Worobey *et al*. indicated that a substantial amount of HIV-1 group M diversity was present in Western Central Africa by 1960, i.e. long before the AIDS pandemic was recognised. This is broadly in line with previous work which used root-to-tip divergence plots to date the TMRCA of HIV-1 group M to the 1930s (Korber *et al.* 2000).

Evidence implicating Kinshasa, the capital of the DRC, as the origin of the HIV-1 group M pandemic is outlined by Sharp and Hahn (Sharp and Hahn 2008; Sharp and Hahn 2011). This evidence rests partly on the circulation of all HIV-1 group M subtypes (except subtype B, which is extremely rare in Africa) in Kinshasa, and the fact that a greater diversity of group M sequences has been observed in Kinshasa than in any other location (Vidal *et al.* 2000; Rambaut *et al.* 2001). In addition, the oldest HIV-1 group M strains, ZR59 and DRC60, were sampled in Kinshasa. Both census data and coalescent-based demographic analyses reveal a population expansion in western central Africa over the period of time in which the HIV-1 group M virus came to prominence (Worobey *et al.* 2008). Finally, Cameroon has been identified as the probable location for a cross-species transmission of SIV from chimpanzees to humans, and is connected to Kinshasa via the Congo River, which was traditionally used as a trade route between the two regions (Sharp and Hahn 2008).

Previous analyses of early HIV sequences from the DRC have revealed that isolates from a number of patients fall in different phylogenetic positions in trees constructed separately for sequences from the *gag* and *env* regions (Robertson *et al.* 1995; Vidal *et al.* 2000; Kalish *et al.* 2004; Yang *et al.* 2005a). Vidal *et al.* (2000) analysed HIV-1 group M sequences sampled in the DRC in 1997, finding discordant *gag* (p24) and *env* (V3-V5) subtypes in 29% of samples and providing evidence that most subtypes were involved in inter-subtype recombination. Subsequently, Kalish *et al.* (2004) found over 25% of sampled infected hospital workers in Kinshasa in the mid-1980s to have discordant *gag* (p17) and *env* (gp41) subtypes. A study of high-risk individuals in Kinshasa suggested that the proportion of individuals infected with inter-subtype recombinant viruses had not changed substantially between 1985 and 2000 (Yang *et al.* 2005a). However, whilst the prevalence of inter-subtype recombinant HIV viruses can be estimated from sequence alignments, little is known about the frequency with which such viruses arise *in vivo* (Chin *et al.* 2005) and no estimates are available for the rate at which they contribute to HIV diversity at the inter-host (population) level.

In Chapter 5, I investigated the use of discrete trait mapping methods for quantifying reassortment between different NA subtypes in H7 avian influenza. This was the first time that 'Markov jumps' counting (see Section 2.10.3) had been used for such a purpose. I showed that Markov jumps counting offered an alternative approach to that of Lycett *et al*. (2012), who used the methods of Lemey *et al*. (2009) to estimate relative instantaneous transition rates between HA and NA subtypes in order to study swine influenza reassortment. In this chapter, I use Bayesian discrete ancestral trait mapping methods in BEAST to quantify inter-subtype recombination in population-level phylogenies for HIV-1 group M, by comparing trees for protein-coding sequences at different ends of the HIV-1 genome. I use Markov jumps counting to calculate the Number of Excess ancestral Subtype Transitions (NEST) required to map individuals' viral subtypes for sequences from one end of the genome (*gag* p17 region) onto the tree for the opposite end of the genome (*env* gp41 region). This quantity can then be rescaled to estimate the rate at which such events have arisen across the phylogenies as a result of inter-subtype recombination. The method is applied to the Kinshasa HIV-1 group M dataset of Kalish *et al*. (2004), which provides an unparalleled opportunity to investigate recombination in a population where multiple subtypes were freely mixing.

6.4 Methods

6.4.1 Data

The HIV sequences used in this study were previously published and analysed by Kalish *et al*. (2004) (Appendix D, Table D1). Serum samples were taken as part of a cross-sectional study of hospital workers at the Mama Yemo Hospital in Kinshasa during the Projet SIDA surveillance program, which operated in the DRC between

1984 and 1991. HIV-1 sequences were obtained via consensus sequencing of PCR-amplified RNA from serum, from samples where sufficient quantities of serum were available. All sequences corresponded to samples obtained between 1984 and 1986.

All HIV-1 sequences for the *gag* p17 and *env* gp41 regions published by Kalish *et al*. (2004) were downloaded from GenBank. The p17 and gp41 sequences from the same individual could be matched by the patient identifier in the sequence label. Datasets were created containing only sequences from persons (a total of 57) for whom both a gp41 and p17 sequence was available. The sequences were aligned manually using BioEdit (Hall 1999) and the alignments were 429 base pairs and 369 base pairs in length for p17 and gp41 respectively. In order to assign subtypes to the p17 and gp41 sequences, reference sequences from the Los Alamos HIV Database (www.hiv.lanl.gov) were downloaded for each subtype, as well as for CRFs 01 and 02. Chimpanzee (CPZ.CM.1998.CAM3.AF115393) gp41 and p17 sequences which fall basal to HIV groups M and N (Hahn *et al.* 2000), and reference sequences for group N and group O viruses (Ref.N.CM.95.YBF30.AJ006022 and Ref.O.BE.87.ANT70.L20587), were downloaded for use as outgroups in the preliminary (neighbor-joining and maximum likelihood) phylogenetic analysis.

6.4.2 Subtyping and preliminary phylogenetic analysis

Phylogenetic analyses were conducted to assign subtypes to the p17 and gp41 sequences, since some uncertainty had been reported in the neighbor-joining analysis of Kalish *et al*. (2004). Maximum likelihood phylogenetic trees for p17 and gp41 were constructed in PhyML (Guindon *et al.* 2010), with 1000 bootstrap replicates (Appendix D, Figure D1 and Figure D2). A general time reversible model (Tavaré 1986) of nucleotide substitution was implemented, with gamma distributed rate heterogeneity across sites and four rate categories. The effect of using different outgroups (group N, O or chimpanzee sequences), or a midpoint rooted tree, was considered. Sequences were classified as a particular subtype if they belonged to a clade containing a reference sequence of that subtype, and no reference sequences of any other subtype. Sequences which were basal to clades containing two or more subtypes were labelled as ‘unclassified’.

Preliminary Bayesian phylogenetic analysis was carried out using BEAST (Drummond and Rambaut 2007), to confirm the subtyping of sequences from the ML analysis and determine the molecular clock model providing the best fit to the data. A relaxed demographic prior (Bayesian skyline with 5 bins) (Drummond *et al.* 2005) was implemented, and the SRD06 nucleotide substitution model (Shapiro *et al.* 2006) was used. Bayes factor testing indicated that an uncorrelated relaxed lognormal clock model was preferred to a strict molecular clock. Since precise sample-date information was not available for the sequences, the mean substitution rate for the uncorrelated lognormal relaxed clock model was fixed to 1, returning branch lengths in units of substitutions per site.

Markov chain Monte Carlo (MCMC) sampling took place every 10,000 generations over a period of 100 million generations in all BEAST runs, with a burnin period of 10 million generations. The chain traces were inspected in the Tracer software (Drummond and Rambaut 2007) (available from http://beast.bio.ed.ac.uk/Tracer) to indicate whether stationarity had been achieved, and multiple runs were compared for all analyses. Effective sample sizes (ESS) were greater than 200 for all parameters estimated. Bayesian skyline plots were constructed using the Tracer software in order to visualise changes in relative genetic diversity between the root and tips of the trees for both the p17 and gp41 datasets.

6.4.3 Within-gene recombination analysis

In order to investigate whether recombination had taken place within the gene fragments encoding the gp41 and p17 proteins, two different analyses were undertaken. Firstly, a single breakpoint analysis (Kosakovsky Pond *et al.* 2006b; Kosakovsky Pond *et al.* 2006a) (Section 2.13) was performed on the individual p17 and gp41 alignments, using an HKY model of nucleotide substitution and allowing gamma-distributed rate heterogeneity across sites with 4 rate classes. These analyses were carried out using HyPhy (Kosakovsky Pond *et al.* 2005) on the DataMonkey web-server (Delport *et al.* 2010) (www.datamonkey.org). Secondly, a sliding window analysis was conducted upon the individual p17 and gp41 alignments, using the TreeOrderScan procedure in the SSE software (Simmonds 2012) (Section 2.13), to assess consistency in phylogenetic clustering along the length of the alignment. Neighbor-joining trees were constructed, with the chimpanzee SIV sequence CPZ.CM.1998.CAM3.AF115393 used as an outgroup. Since this procedure requires the user to choose the length of fragments of the alignment upon which to construct phylogenies, as well as the intervals at which to construct phylogenies, different combinations of settings were investigated: alignment fragments of length 100 and 150 nucleotides, at intervals of both 25 and 50 nucleotides. In the visual output, sequences were coloured according to their assigned subtype from the ML analysis to enable any evidence of recombination to be classified as intra-subtype (seen as exchange of phylogenetic position between lines of the same colour) or inter-subtype (which would be observed as exchange of phylogenetic position between lines of different colours).

6.4.4 Inter-subtype recombination analysis

Sequences in the p17 and gp41 alignments were labelled according to both the p17 and gp41 subtypes for that individual, and inter-subtype recombinant viruses were identified by a discrepancy between p17 and gp41 subtypes from an individual. Discrete ancestral trait mapping was performed in BEAST to infer ancestral subtypes along the posterior phylogeny samples, starting with the subtypes at the tips of the tree and modelling transitions between ancestral subtypes as an asymmetric continuous-time Markov process, using the implementation of Lemey *et al*. (2009). In order to quantify the amount of inter-subtype recombination between opposite ends of the genome, both the p17 and gp41 subtypes were independently mapped onto the p17 and gp41 phylogeny samples. The same BEAST MCMC settings were used as described for the preliminary BEAST analysis. The procedure for quantifying inter-subtype recombination is outlined in Figure 6.2.

The number of ancestral p17 and gp41 subtype transitions along each BEAST phylogeny sample was recorded by manually editing the BEAST xml file to employ the ‘Markov jumps’ method (Minin and Suchard 2008a; Minin and Suchard 2008b; O’Brien *et al*. 2009; Talbi *et al*. 2009) (see Section 2.10.3) for counting discrete trait transitions across the phylogeny samples. Four sets of ancestral subtype transition counts were obtained from the Markov jumps analysis:

(i) the number of p17 subtype transitions on the p17 trees (denoted p17_p17)
(ii) the number of gp41 subtype transitions on the p17 trees (gp41_p17)
(iii) the number of p17 subtype transitions on the gp41 trees (p17_gp41)
(iv) the number of gp41 subtype transitions on the gp41 trees (gp41_gp41).

An example visualisation of the number of transitions along p17 and gp41 phylogenies is provided in Appendix D (Figure D3). The Number of Excess ancestral Subtype Transitions (NEST) required to map subtypes onto the phylogeny for the wrong gene (p17_gp41 or gp41_p17) compared to onto the phylogeny for the correct gene (p17_p17 or gp41_gp41) was calculated for 9,000 randomly paired gp41 and p17 posterior phylogeny samples. 95% highest posterior density (HPD) intervals were calculated for the NEST across the paired phylogeny samples.

In the absence of inter-subtype recombination, the number of ancestral subtype changes required to map individuals’ subtypes onto a phylogeny for the correct gene (e.g. p17_p17) should be equal to the number of ancestral subtype changes required to map the same subtypes onto the phylogeny for the other end of the genome (e.g. mapping patients’ p17 subtypes onto the gp41 tree). This is because, with no inter-subtype recombination, the structure of the gp41 and p17 trees with regard to branching into subtypes should be essentially the same, and in this case the NEST would be centred on zero. The effect of inter-subtype recombination is to ‘shuffle’ the subtypes at the tips, meaning that in general more ancestral subtype transitions would take place along the tree when the subtypes from one end of the genome were mapped onto a phylogeny constructed for the other end of the genome, compared with mapping subtypes onto the phylogeny for the correct gene. Since inter-subtype recombination events create excess ancestral subtype transitions along phylogenies constructed from the opposite end of the genome, the NEST allows the amount of inter-subtype recombination which can be detected between phylogenies to be quantified.

NEST estimates were obtained by mapping p17 subtypes onto the p17 and gp41 phylogenies, as well as mapping gp41 subtypes onto gp41 and p17 phylogenies. The NEST estimates were then rescaled to estimate the rate (per lineage, per year) at which excess subtype transitions, arising from inter-subtype recombination events, occurred. For each pair of phylogeny samples, the NEST was divided by the sum of the branch lengths (in units of substitutions per site) of the tree from the opposite end of the genome to the subtypes being mapped (e.g. gp41 tree when mapping p17 subtypes) then multiplied by an estimate of the rate of HIV-1 nucleotide substitution of 2.47×10^{-3} substitutions/site/year (Worobey *et al.* 2008).

As reported from the maximum likelihood analysis, as well as by previous studies (Rambaut *et al.* 2001; Kalish *et al.* 2004), it is not always possible to unambiguously assign subtypes to HIV-1 group M sequences from the Democratic Republic of the Congo. In particular, some sequences appear to fall basal to clades of more than one subtype. The potential for difficulty in assigning subtypes to have introduced error into the analysis was investigated by repeating the analysis using the alternative subtype labellings for sequences which were difficult to classify (Appendix D, Table D1). In addition, analyses were performed where sequences were labelled by the clade to which they belonged at a cut-off defined at the root of the subtype A clade in the maximum clade credibility (MCC) tree. Defining a cut-off at the root of the subtype A clade resulted in 10 clades being defined on the gp41 tree, and 7 or 10 clades on the p17 tree (Appendix D, Figure D4 and Figure D5) i.e. similar to the number of subtypes identified in the maximum likelihood analysis. The excess number of transitions required to map the p17 labels onto the gp41 trees, compared to mapping them onto the p17 trees, was then compared in the same manner as using the subtype labels.

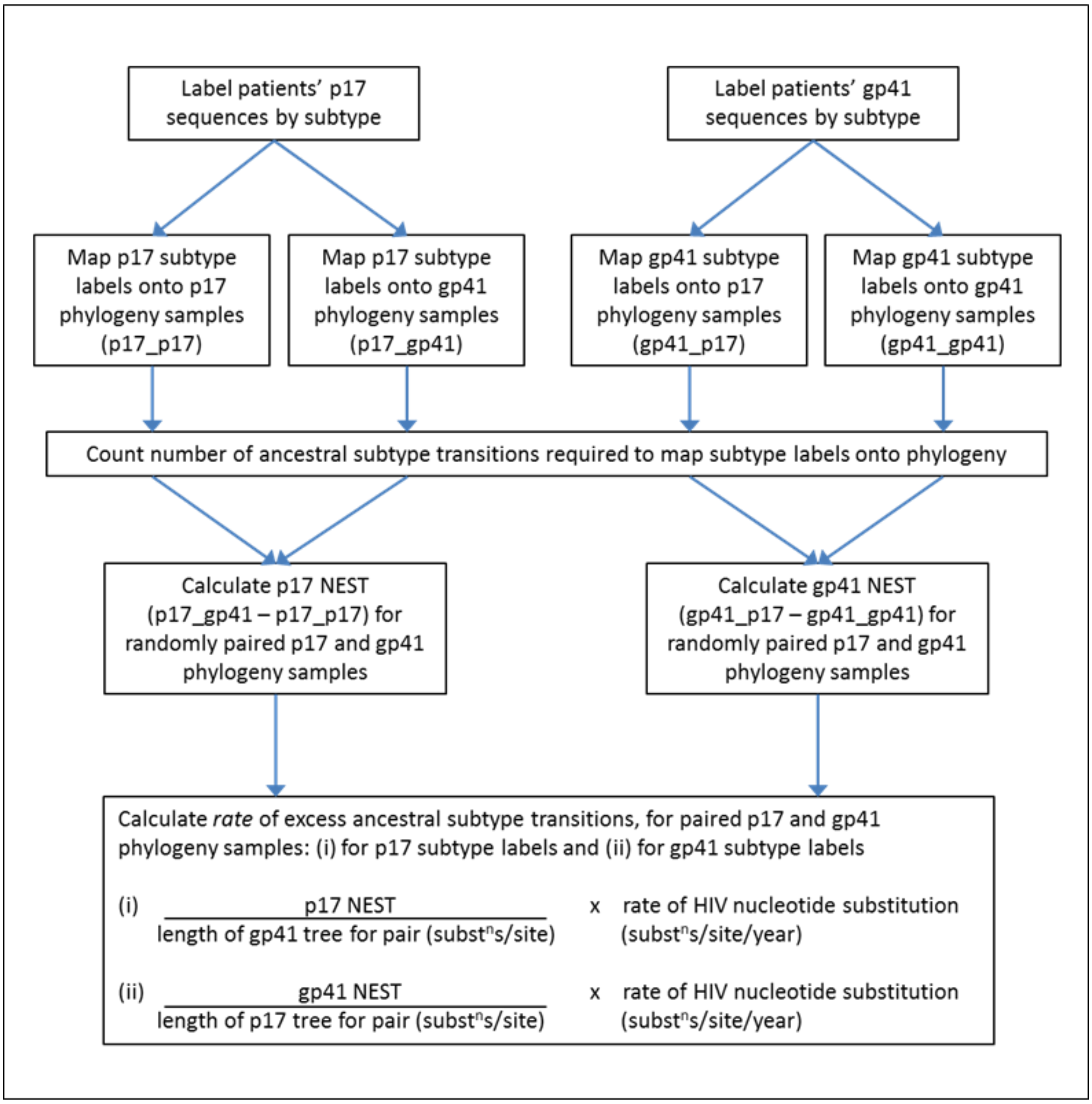

Label patients' p17 sequences by subtype
Label patients' gp41 sequences by subtype
Map p17 subtype labels onto p17 phylogeny samples (p17_p17)
Map p17 subtype labels onto gp41 phylogeny samples (p17_gp41)
Map gp41 subtype labels onto p17 phylogeny samples (gp41_p17)
Map gp41 subtype labels onto gp41 phylogeny samples (gp41_gp41)
Count number of ancestral subtype transitions required to map subtype labels onto phylogeny
Calculate p17 NEST (p17_gp41 – p17_p17) for randomly paired p17 and gp41 phylogeny samples
Calculate gp41 NEST (gp41_p17 – gp41_gp41) for randomly paired p17 and gp41 phylogeny samples
Calculate *rate* of excess ancestral subtype transitions, for paired p17 and gp41 phylogeny samples: (i) for p17 subtype labels and (ii) for gp41 subtype labels
(i) p17 NEST / length of gp41 tree for pair (subst^{n}s/site) x rate of HIV nucleotide substitution (subst^{n}s/site/year)
(ii) gp41 NEST / length of p17 tree for pair (subst^{n}s/site) x rate of HIV nucleotide substitution (subst^{n}s/site/year)

Figure 6.2
Process for inferring the rate of excess ancestral subtype transitions resulting from inter-subtype recombination. The number of excess of p17 or gp41 ancestral subtype transitions (NEST) required to map the subtype labels onto the phylogeny for the wrong gene compared to onto the phylogeny for the correct gene was calculated. Rescaling by the length of the tree and a time-scaled rate of nucleotide substitution was then carried out to calculate the rate at which excess ancestral subtype transitions occurred, in order to provide a measure of the rate of inter-subtype recombination on an explicit timescale.

6.5 Results

6.5.1 Dataset composition

For each individual studied, the HIV-1 gp41 and p17 sequences were labelled in the form 'p17 subtype_gp41 subtype_ two letter country code_patient ID number'. 'Potentially pure' viruses, where an individual's HIV-1 p17 and gp41 sequences were of the same subtype, were present in the Kinshasa dataset for all subtypes except H, which is globally rare, and B, which does not appear amongst African sequences from this time (Vermund and Leigh Brown 2011). Discordant p17 and gp41subtypes were found in 26% of individuals (Figure 6.3), in line with previous analysis of this dataset (Kalish *et al.* 2004). Nine different discordant p17 and gp41 subtype combinations were present in the dataset (Table 6.1) and subtypes A, D, F, G, H and J were involved in inter-subtype recombination events. Subtype A, which was the most frequently isolated pure subtype, was also the most commonly represented subtype amongst the recombinant viruses, with 11 out of the 15 (73%) recombinant viruses having a p17 or gp41 sequence of subtype A. The most frequently occurring recombinant virus was of type A_G (3 out of 15, i.e. 20%, of the recombinant viruses), and there was also one G_A inter-subtype recombinant virus. Additionally, three viruses were labelled as recombinants since their p17 sequences were of subtype A whilst their gp41 sequences formed a clade of their own, clustering in the maximum likelihood trees with reference sequences of type CRF 01.

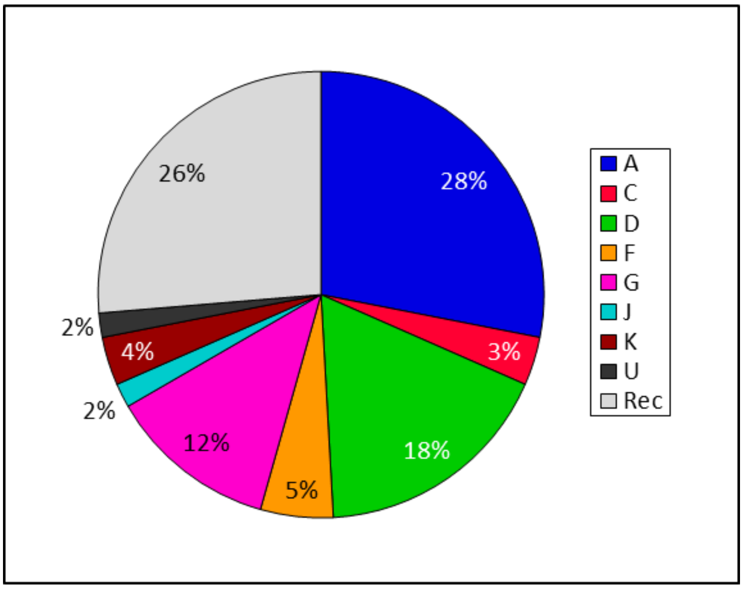


Figure 6.3
Subtype distribution of HIV-1 group M in Kinshasa. The gp41 and p17 regions of HIV-1 group M were sequenced for 57 patients by Kalish *et al.* (2004). The percentage of patients infected with 'potentially pure' viruses (i.e. with the same gp41 and p17 subtype) of a given subtype on the basis of maximum likelihood phylogenetic analysis is reported. 26% of the viruses were classified as recombinant ('Rec.') on the basis of different subtypes having been assigned to the p17 and gp41 regions.

p17 subtype	gp41 subtype	Frequency
A	G	3
A	H	1
A	J	2
A	CRF01	3
D	F	2
D	G	1
G	A	1
H	A	1
J	U	1

Table 6.1
Frequency of recombinant types. The p17 and gp41 subtypes of discordant sequences, and their frequency of occurrence in the dataset, were reported. Note that 'U' denotes an unclassifiable sequence and CRF01 denotes the circulating recombinant form previously known as subtype E, which forms a distinct clade in the gp41 region.

Ancestral trait mapping of p17 and gp41 subtypes was performed upon sets of BEAST phylogenies for the p17 and gp41 regions, and limited evidence for clustering of the recombinant viruses was observed in the maximum clade credibility (MCC) trees (Figure 6.4). The two gp41 sequences from D_F viruses clustered together in the gp41 tree (posterior probability $p = 0.902$), and their p17 sequences were also sister lineages ($p = 0.686$). In the p17 tree, two of the three A_CRF01 viruses clustered together in the p17 tree ($p = 0.917$) and also clustered in the gp41 tree ($p = 0.998$). Since clustered HIV sequences are often considered to be epidemiologically linked (Lewis *et al.* 2008), these clusters may represent a single inter-subtype recombination event, followed by transmission of the recombinant virus. Clusters of inter-subtype recombinant viruses arising from a single recombination event, followed by transmission of the recombinant virus, would only incur one additional ancestral subtype transition in our method for quantifying recombination, whereas multiple independent inter-subtype recombinations across the tree would require further additional transitions.

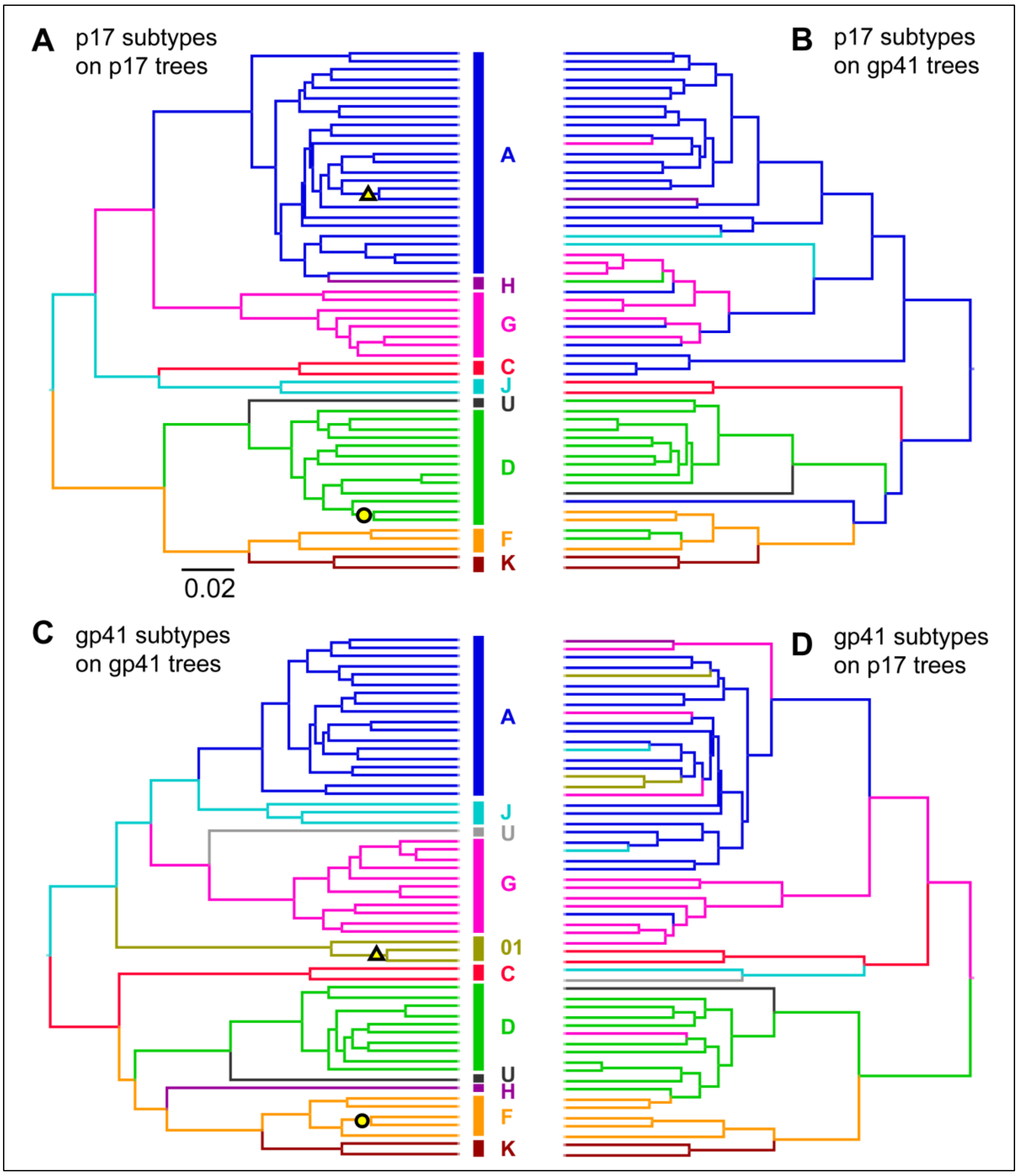


Figure 6.4
Maximum clade credibility trees for Kinshasa 1984 dataset, coloured by ancestral subtype. Maximum clade credibility (MCC) trees were constructed using BEAST. Branches were coloured according to inferred ancestral subtypes, mapping (A) patients' p17 subtypes onto the p17 tree samples; (B) patients' gp41 subtypes onto the gp41 trees; (C) gp41 subtypes onto the gp41 trees and (D) gp41 subtypes onto the p17 trees. The number of p17 and gp41 subtype transitions (Markov jumps) across the tree was recorded for each posterior phylogeny sample. Clustering of recombinant sequences can be observed in the MCC trees for two D_F patients (marked with circles; posterior probabilities of being sister lineages = 0.686 and 0.902 in the p17 and gp41 trees respectively) and two A_CRF01 patients (marked with triangles; posterior probabilities of being sister lineages = 0.917 and 0.998 in the p17 and gp41 trees respectively). Branch lengths are in units of substitutions per site.

6.5.2 Within-gene recombination analysis

No evidence of recombination was detected within the individual gp41 or p17 alignments by the single breakpoint analysis under either the BIC or corrected AIC (AICc). A sliding-window analysis using TreeOrderScan indicated that any exchange in phylogenetic position along the individual alignment occurred between sequences to which the same subtype had been assigned. This could be observed in Figure 6.5a and Figure 6.6a, where the lines representing sequences were coloured according to subtype assigned from the ML analysis, and crossing-over only occurs between lines of the same colour. Such patterns indicated that mosaic gp41 or p17 sequences with sections belonging to more than one subtype were not present. In the sliding window analysis of the gp41 alignment using fragments of length 150 nucleotides at intervals of 50 nucleotides (Figure 6.5b and Figure 6.6b), the sequence which had been labelled ‘unclassified’ appeared to change phylogenetic position along the genome (although not when fragments of length 100 nucleotides were used), perhaps explaining the difficulties experiences with subtyping this sequence against the reference sequences.

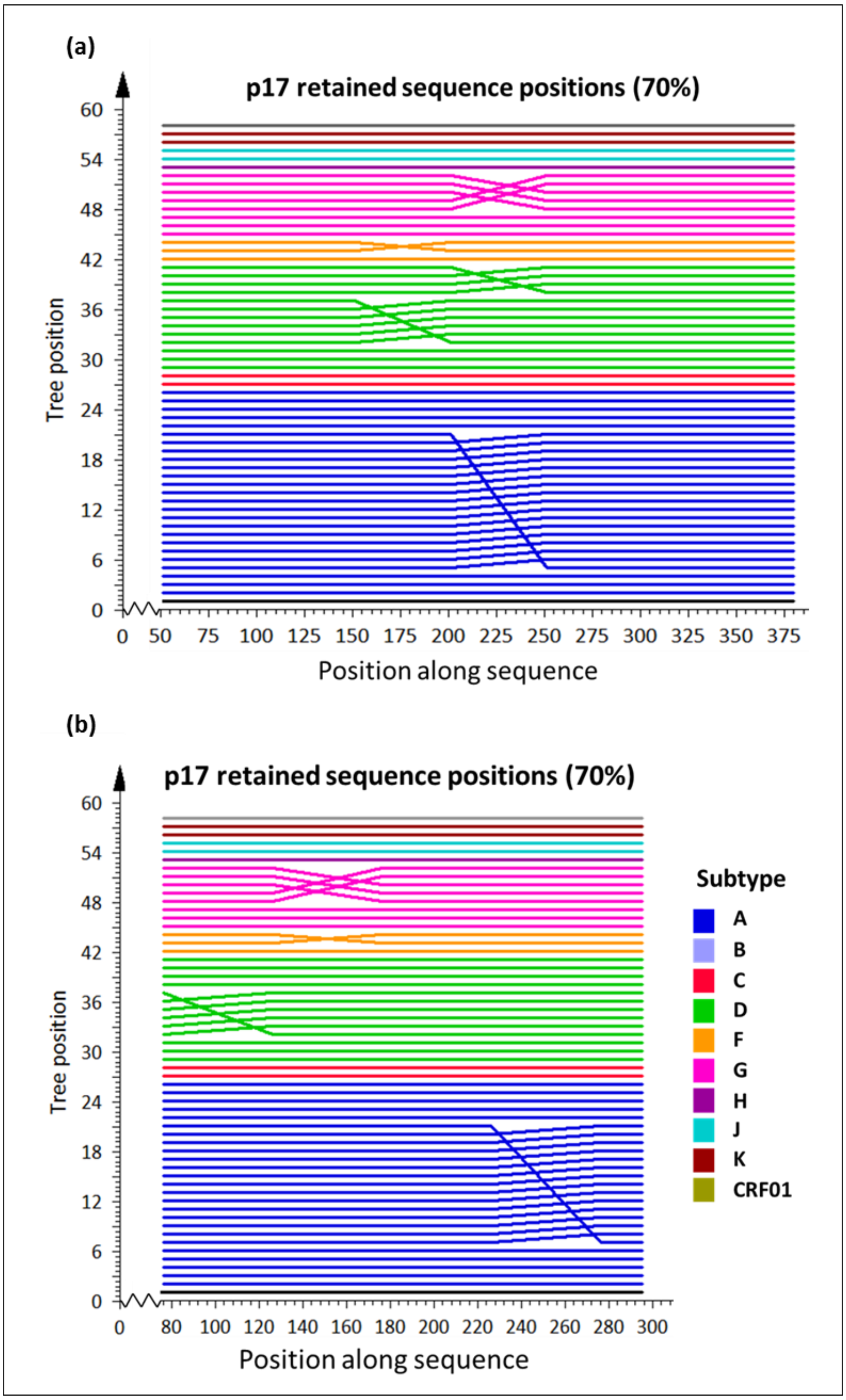


Figure 6.5
Retained sequence positions across the p17 region. A TreeOrder Scan analysis was performed in the SSE software by constructing neighbour-joining trees upon fragments of **(a)** length 100 and **(b)** 150 nucleotides, at intervals of 50 nucleotides. Any detectable change in phylogenetic position, based upon support for clades with a bootstrap value of greater than 70%, occurred amongst sequences to which the same subtype had been assigned (represented by lines of the same colour). There was therefore no evidence of inter-subtype recombination within the p17 gene. Analyses were also performed on fragments of length 100 and 150, at intervals of 25 nucleotides, and no crossing over between lines of different colours was observed (not shown).

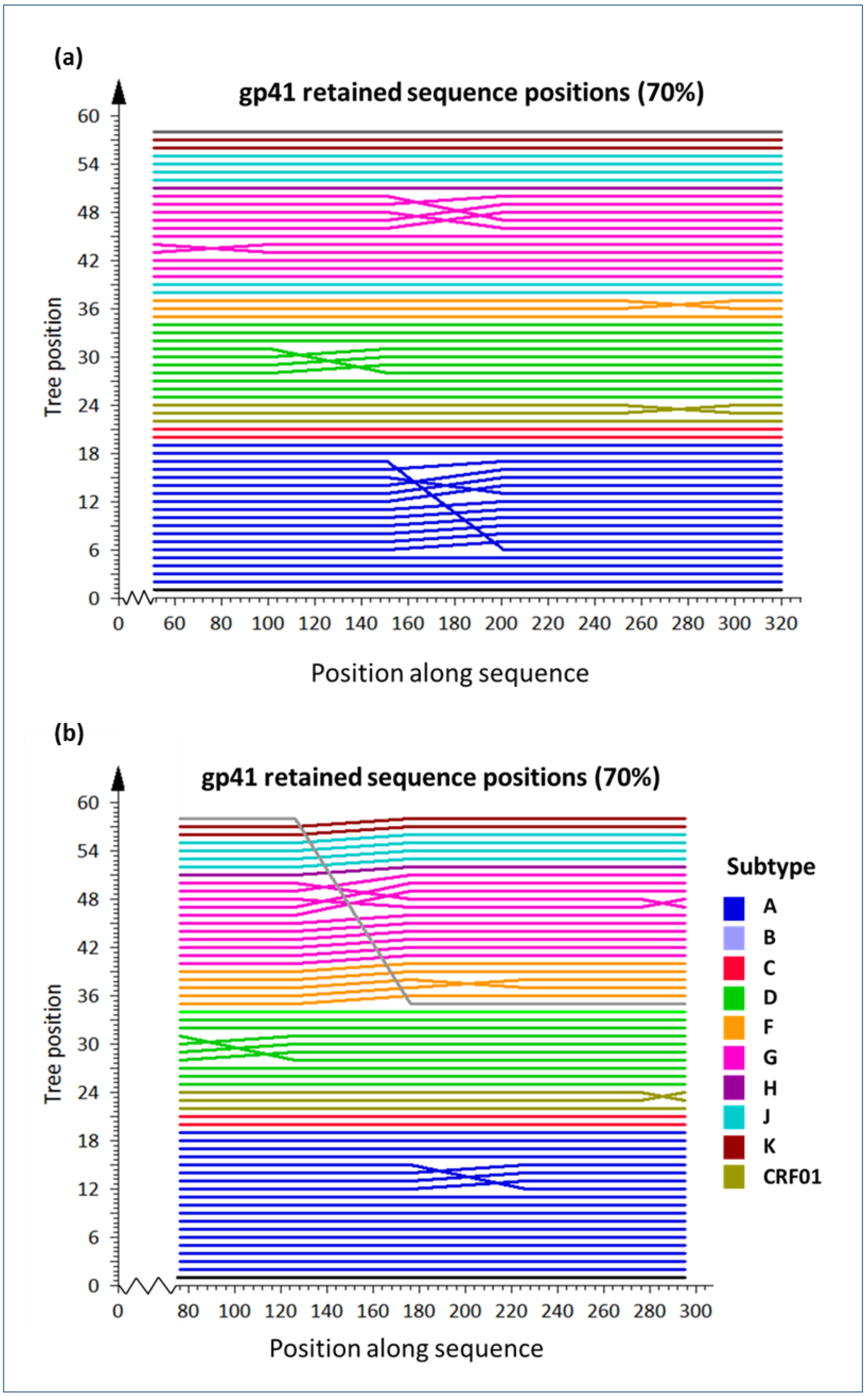


Figure 6.6
Retained sequence positions across the gp41 region. A TreeOrder Scan analysis was performed in the SSE software by constructing neighbour-joining trees upon fragments of **(a)** length 100 and **(b)** 150 nucleotides, at intervals of 50 nucleotides. In (a), detectable change in phylogenetic position occurred amongst sequences to which the same subtype had been assigned. However, when fragments of length 150 nucleotides rather than 100 nucleotides were used, the sequence represented by the grey line at top left of plot appeared to change phylogenetic position across the gp41 region, perhaps explaining the difficulty in subtyping this sequence from the maximum likelihood trees containing reference sequences, which had led to it being labelled ‘unclassified’. Similar patterns were observed when the analysis was re-run using intervals of 25 nucleotides in length (not shown).

6.5.3 Quantifying inter-subtype recombination

Discrete trait mapping methods in BEAST (Lemey *et al.* 2009) have previously been used for phylogeographic analysis of viral sequence data (e.g. Raghwani *et al*. (2011)), but can also be used to infer other ancestral character traits onto phylogenies. If there are *k* different states at the tips of a phylogeny, the minimum number of ancestral state transitions observed across the phylogeny would be *k*-1. When HIV-1 group M subtypes were mapped onto the BEAST phylogeny for the correct gene (i.e. p17 subtypes on the p17 tree, or gp41 subtypes on the gp41 tree), the number of ancestral subtype transitions across the tree lay towards this minimum number (Figure 6.7 and Figure 6.8) and phylogenetic uncertainty accounted for instances where more than the minimum number of transitions was required. A greater number of ancestral subtype transitions were required to map patients' gp41 or p17 subtypes onto phylogeny samples for the other end of the genome (p17 subtypes on the gp41 tree, or gp41 subtypes on the p17 tree), compared to the correct gene, as a result of inter-subtype recombination (Figure 6.7 and Figure 6.8). Each additional ancestral subtype transition could be interpreted as arising from inter-subtype recombination; hence, inter-subtype recombination is detectable from the phylogenies in this way.

Results from the ancestral subtype mapping analyses are reported in Table 6.2. The mean Number of Excess ancestral Subtype Transitions (NEST) required to map the p17 subtypes onto the gp41 phylogeny samples, compared to onto the p17 phylogeny samples, was 10.55 (95% HPD interval = 2, 18). The mean NEST for mapping gp41 subtypes onto p17 and gp41 phylogeny samples was 12.18 (95% HPD interval = 5, 20). When the NEST was re-scaled as described in Section 6.4.4 and Figure 6.2, the rate at which excess ancestral substitutions arose was estimated to be 6.93x10^{-3} per lineage per year (95% HPD interval = 2.39x10^{-3}, 1.30x10^{-2}) using p17 subtype labels, and 8.11x10^{-3} per lineage per year (95% HPD interval = 3.11x10^{-3}, 1.41x10^{-2}) using gp41 subtype labels.

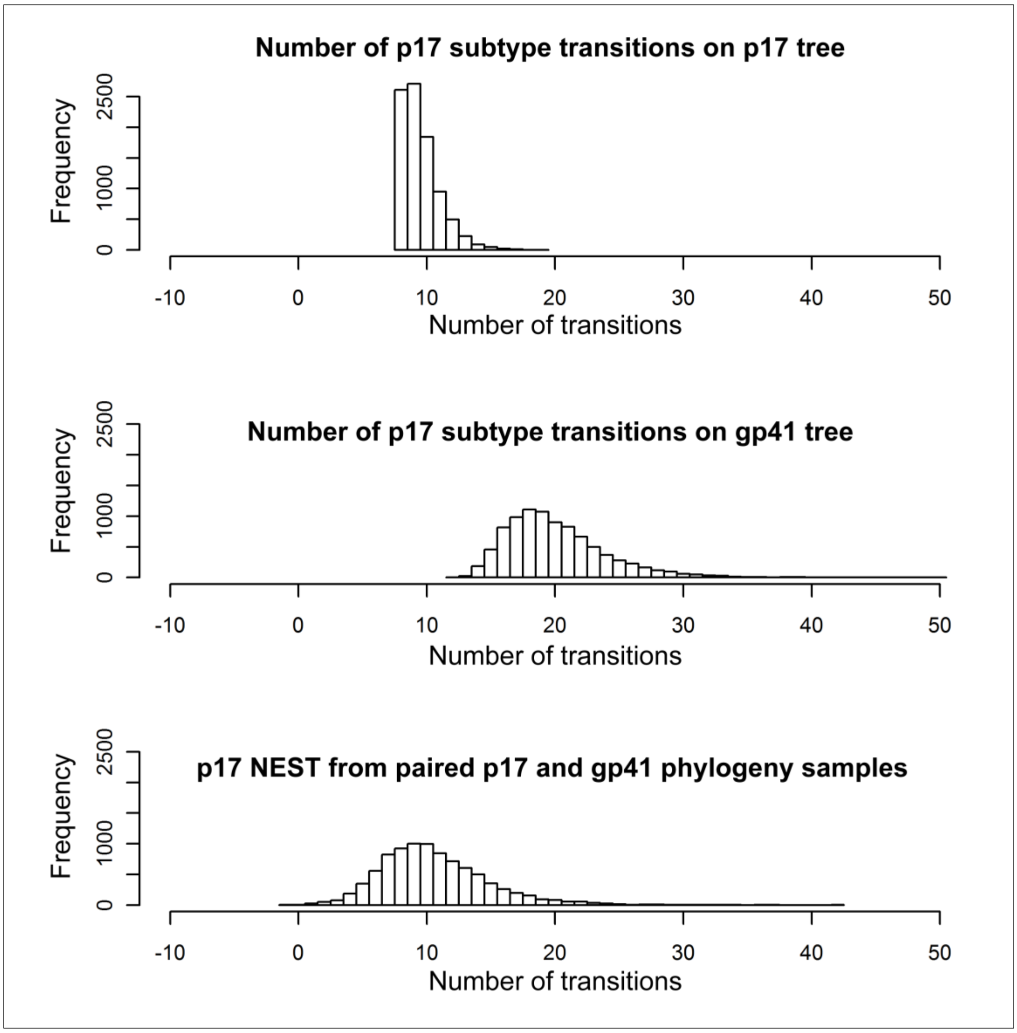


Figure 6.7
Number of inferred p17 ancestral subtype changes across phylogeny samples. The number of ancestral subtype transitions across phylogeny samples was inferred using ancestral trait mapping in BEAST, mapping p17 subtypes onto the p17 and gp41 phylogeny samples. The number of excess subtype transitions (NEST) required to map the ancestral p17 subtypes onto the phylogeny for the ‘wrong’ gene, compared to mapping them onto the correct phylogeny, was calculated across paired phylogeny samples. The histograms represent the number of subtype transitions across 9,000 post-burnin samples of phylogenies.

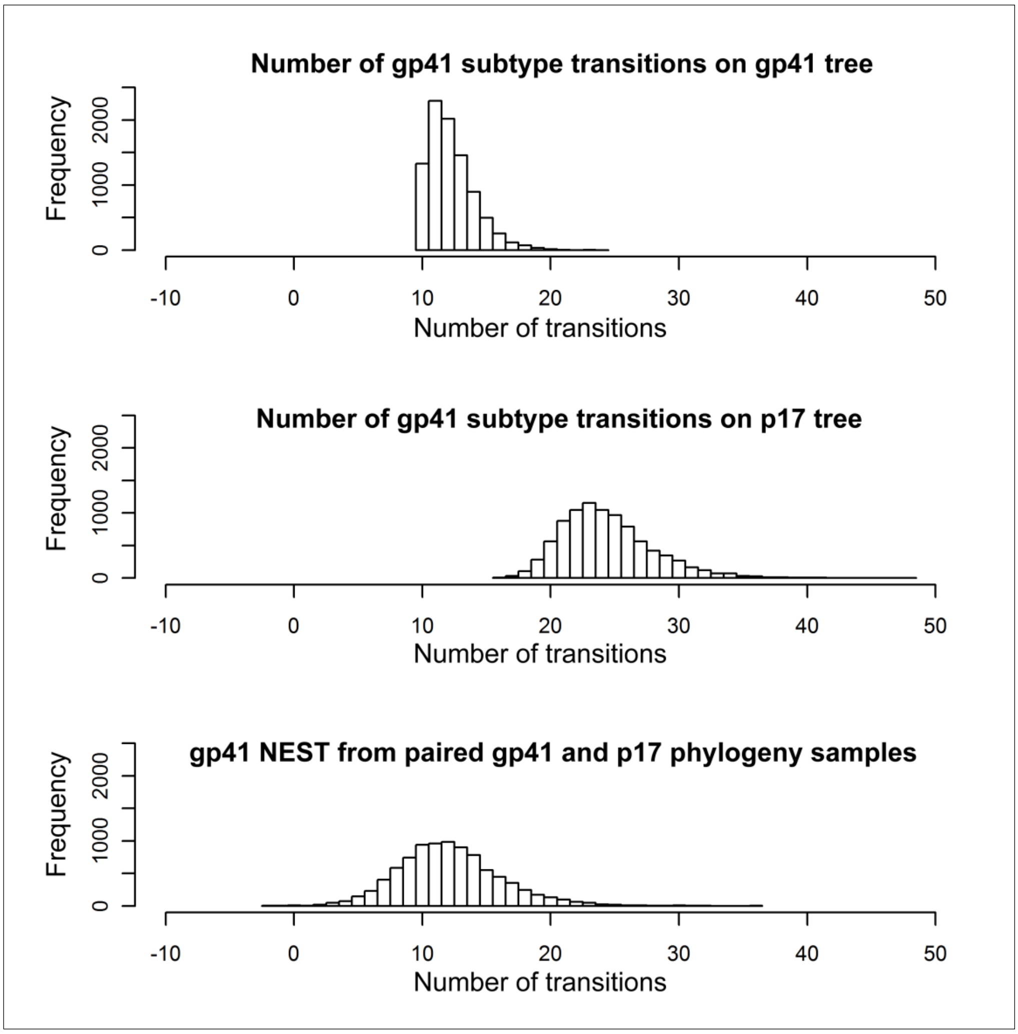


Figure 6.8
Number of inferred gp41 ancestral subtype changes across phylogeny samples. The number of ancestral subtype transitions across phylogeny samples was inferred using ancestral trait mapping in BEAST, mapping gp41 subtypes onto the p17 and gp41 phylogeny samples. The number of excess subtype transitions (NEST) required to map the ancestral gp41 subtypes onto the phylogeny for the 'wrong' gene, compared to mapping them onto the correct phylogeny, was calculated across paired phylogeny samples. The histograms represent the number of jumps across 9,000 post-burnin samples of phylogenies.

Subtype labels	Transitions on p17 tree	Transitions on gp41 tree	NEST	Excess transitions per year
gp41	24.44 (18, 31)	12.25 (10, 16)	12.18 (5, 20)	8.11x10^{-3} (3.11x10^{-3}, 1.41x10^{-2})
gp41_labels_i	22.59 (16, 29)	9.79 (8,13)	12.80 (5,20)	8.52x10^{-3} (3.48x10^{-3}, 1.43x10^{-2})
gp41_labels_ii	23.99 (17, 31)	11.10 (9, 14)	12.89 (5, 21)	8.60x10^{-3} (3.18x10^{-3}, 1.46x10^{-2})
gp41_rootA	23.10 (17, 29)	10.93 (9, 14)	12.17 (5, 19)	8.20x10^{-3} (3.22x10^{-3}, 1.35x10^{-2})
p17	9.50 (8, 12)	20.05 (13, 27)	10.55 (2, 18)	6.93x10^{-3} (2.39x10^{-3}, 1.30x10^{-2})
p17_labels_i	8.24 (7, 11)	18.86 (13, 26)	10.61 (3, 18)	7.02x10^{-3} (2.34x10^{-3}, 1.28x10^{-2})
p17_labels_ii	9.42 (8, 12)	20.11 (14, 27)	10.69 (4, 19)	7.07x10^{-3} (2.29x10^{-3}, 1.29x10^{-2})
p17_rootA_v1	7.02 (6, 9)	18.15 (13, 26)	11.13 (4, 19)	7.32x10^{-3} (2.80x10^{-3}, 1.32x10^{-2})
p17_rootA_v2	22.32 (15, 30)	10.81 (9, 14)	11.50 (3, 20)	7.56x10^{-3} (2.35x10^{-3}, 1.39x10^{-2})

Table 6.2
Results from discrete ancestral trait mapping of subtypes onto p17 and gp41 trees. The number of ancestral subtype transitions required to map p17 and gp41 subtype labels onto BEAST phylogeny samples constructed from p17 and gp41 sequences was reported. The number of excess subtype transitions (NEST) required to map the subtypes onto the phylogeny for the correct gene, compared to onto the phylogeny from the other end of the genome, was rescaled by the sum of the branch lengths of the phylogeny and the rate of HIV-1 nucleotide substitution in order to obtain a rate on an explicit timescale. Results are shown for alternative labellings of ambiguous sequences ('labels_i and ii'). Results for 'gp41_rootA', 'p17rootA_v1' and 'p17rootA_v2' refer to the analyses where gp41 and p17 labels were assigned from the root of the subtype A clade (see Figure D4 and Figure D5). Numbers in brackets are the 95% HPD limits.

The substantial overlap of the HPD intervals for the NEST, and for the excess subtype transition rate estimates, indicated that the estimates obtained using p17 and gp41 subtypes were not significantly different. Slight differences may have arisen because different numbers of gp41 subtypes and p17 subtypes (9 and 11 respectively) were present in the dataset. HPD intervals were similar when alterative subtype labellings were used for sequences which were difficult to classify, and when clades were defined from a predetermined cut-off point along the tree, indicating that the estimates were robust to potential errors in subtyping of sequences.

6.6 Discussion

Understanding recombination as an ancestral process is important for unravelling the evolutionary history of HIV and explaining the pattern of HIV diversity (Abecasis *et al.* 2007). In addition, recombination can confound phylogenetic analyses which assume that a single evolutionary tree applies to the whole of an alignment (Posada and Crandall 2002), leading to false positives when detecting sites under positive selection (Anisimova *et al.* 2003; Shriner *et al.* 2003) and affecting estimates of divergence dates (Schierup and Hein 2000; Worobey 2001), or at least increasing the variance of such estimates (Lemey *et al.* 2004).

Although previous studies have investigated crossover rates *in vitro*, such studies do not measure the rate at which inter-subtype recombination contributes to HIV diversity at the inter-host phylogenetic level, which is a more complex, composite process. For an inter-subtype recombination event to be detected from a population-level phylogeny, an individual must firstly be infected with viruses of more than one subtype, an inter-subtype recombination must take place and the resulting recombinant virus must be viable and become the dominant strain within an individual. Whilst procedures have been developed for detecting recombination on the basis of phylogenetic discordance (e.g. Posada and Crandall (2001)), methods for quantifying recombination across phylogenies are lacking (Philippe *et al.* 2005). It has therefore been difficult to compare the rate of recombination and other evolutionary processes, such as nucleotide substitution, which contribute to the observed diversity of HIV-1 group M at the population level.

The purpose of this study was to quantify historical inter-subtype recombination events across HIV-1 group M phylogenies, instead of simply calculating the percentage of recombinant sequences in the dataset, as had previously been carried out for this, and other, datasets (e.g. Robertson *et al*. (1995), Vidal *et al*. (2000), Kalish *et al*. (2004), Yang *et al*. (2005a)). By analysing viral sequence data from Kinshasa, where almost all HIV-1 group M subtypes co-circulate, the observed recombination events could reasonably be assumed to have occurred within this population.

6.6.1 Interpretation of results

The inter-subtype recombination rate estimate for the Kinshasa dataset of $6.93\text{x}10^{-3}$ to $8.11\text{x}10^{-3}$ events per lineage per year was obtained by calculating the rate at which excess ancestral subtype transitions were observed from mapping individuals' subtypes from one end of the genome onto phylogenies for the other end of the genome. The rate estimate could be compared to other evolutionary processes, such as the HIV-1 nucleotide substitution rate, which has previously been estimated as $2.47\text{x}10^{-3}$ substitutions per site, per year (Worobey *et al.* 2008). The HIV-1 genome is 9,700 base pairs in length and thus approximately 24 ($9,700 * 2.47\text{x}10^{-3}$) nucleotide substitutions would be expected to occur per year across a single genome. Although this study estimates the rate of inter-subtype recombination to be considerably lower than the per-genome nucleotide substitution rate, inter-subtype recombination has far greater potential for instantly generating highly novel HIV-1 group M virus strains than the gradual accumulation of nucleotide substitutions, and poses a significant problem for vaccine design (Burke 1997). Furthermore, it is precisely due to the combination of a rapid mutation rate and relatively infrequent recombination events between viruses of different subtypes that HIV-1 group M inter-subtype recombination may be detected (Awadalla 2003).

The Poisson distribution can be used with the NEST rate estimate to calculate the probability that a lineage evolving for a given period of time would have undergone at least one inter-subtype recombination event (e.g. 1 – *exp*[-6.93x10^{-3} * time-period], or 1-*exp*[-8.11x10^{-3} * time-period]). For example, 6.7-7.8% of lineages evolving for 10 years in the population studied would be expected to undergo inter-subtype recombination, as would 18.8-21.6% of lineages evolving for 30 years. Future work could investigate different ways of scaling the number of excess ancestral subtype transitions, on both time-scaled and non-time-scaled phylogenies, to obtain different quantities which may be interpreted biologically.

The amount of HIV-1 group M inter-subtype recombination observed amongst the Kinshasa sequence data in this chapter must be an under-estimate of the actual amount within the population, since the phylogeny represents only a small sample of the infected population. However, the calculated rate of inter-subtype recombination does account for the total time along the phylogeny, by scaling by the sum of the branch lengths. The use of just two sections of the genome for identifying recombinant viruses means that the inter-subtype recombination rate may be underestimated when multiple crossovers have occurred along the genome. For example, viruses could be designated ‘potentially pure’ by having the same gp41 and p17 subtype, but in fact contain a section derived from a different parental subtype in another region. In fact, viruses denoted either ‘potentially pure’ or inter-subtype recombinant may have complex mosaic genomes with sections derived from more than two parental subtypes, and this study cannot provide a measure of this without sequence data from other genomic regions.

It must also be noted that the hospital workers studied may have been at a higher risk of multiple infection than the general Kinshasa population, due to lack of universal precautions to prevent blood-to-blood transmission through the course of their work. However, the HIV-1 group M prevalence estimate of 3.5% for hospital workers in Kinshasa in 1984 (Kalish *et al.* 2004) is in line with estimates for childbearing women (3.1%) and blood donors (3.1%) in Kinshasa from a study conducted in

1997, which suggested that sero-prevalence of HIV-1 had stabilised in Kinshasa since the 1980s (Mulanga-Kabeya *et al.* 1998). These findings have not suggested that the hospital workers exhibited higher levels of HIV infection than the general population, although data on multiple-infection with different subtypes is not available for a direct comparison between risk groups.

As in Chapter 5, it must be noted that the discrete trait transition models of Lemey *et al*. (2009) assume homogeneity of state frequencies over the tree. Such an assumption may be invalid if the relative frequencies of the different subtypes have changed over the history of the sample. Since the sequences used in this study are some of the earliest HIV-1 group M sequences available, it is not possible to further investigate the validity of the assumption by considering the relative frequency of subtypes at earlier time-points. The prevalence of different subtypes over time may be particularly important for studies where the relative frequency of inter-subtype recombination between different pairs of subtypes was being considered (in contrast to this study, where the overall rate of inter-subtype recombination was of primary interest).

6.6.2 NEST method in the context of phylogenetic and population genetic methods for investigating recombination

Previous phylogenetic approaches for investigating recombination have focused on detecting phylogenetic incongruences (e.g. Nagarajan and Kingsford (2011)) rather than quantifying the amount of recombination observed between phylogenies, or estimating population-level recombination rates. In contrast, NEST is a phylogenetic method which allows recombination to be quantified by comparing phylogenies from different regions of the genome. However, the rate of recombination has also been considered in the field of population genetics (reviewed by Stumpf and McVean (2003)), where one centimorgan (cM) in genetic distance between two loci indicates the probability of a recombination between them in one generation to be 0.01. Population-genetic studies, in particular those concerned with linking the human genome sequence and its genetic map, have expressed recombination rate as a mapping function in units of centimorgans per megabase (cM/Mb).

Future theoretical work and simulation studies could attempt to reconcile the inter-subtype rate estimate obtained using NEST with coalescent models from population genetics which incorporate recombination. The coalescent with recombination for a Wright-Fisher population is a birth-death process whereby lineages merge (coalesce) backwards in time at rate $k(k\text{-}1)/2$, where k is the number of lineages in the genealogy, and new lineages are created by recombination at rate $k\rho/2$ (see Wakeley (2009), p207). The parameter ρ is referred to as the population recombination rate, defined as $\rho=2N_e\,r$ for haploids, where the per-site, per-generation recombination rate r is the probability of recombination occurring during a single round of replication. In models which also allow for mutation (with the population mutation rate defined as $\Theta=2N_e\mu$, where μ is the probability of mutation per site, per generation), it is common to express the recombination rate relative to the mutation rate (r/m) to evaluate the relative frequencies of the two processes. This has a similar rationale to comparing the NEST rate with the rate of nucleotide substitution, as was carried out in Section 6.6.1.

The population recombination rate ρ explicitly depends on the effective population size (N_e), which could make analysis more complex in the presence of the HIV-1 group M population expansion suggested by the Bayesian skyline plots for this data (Appendix D, Figure D6) (Carvajal-Rodriguez *et al.* 2007). In order to relate ρ to the NEST method, datasets of sequences could be simulated under different values of ρ using available software such as RECODON (Arenas and Posada 2007). A NEST analysis could then be performed on these datasets to determine which ρ values yield similar NEST rates to those reported for the Kinshasa dataset in this chapter.

Under the basic Kingman coalescent (see Wakeley (2009)), the ancestry of a sample can be represented by a genealogy (Donnelly and Tavaré 1995). More complex graphs, known as Ancestral Recombination Graphs (ARGs), are required to represent the ancestral history for a population in which recombination has occurred (Hudson

1983; Hudson and Kaplan 1988; Griffiths and Marjoram 1996) (Appendix D, Figure D7). ARGs have been adopted by the phylogenetic networks community (see Bloomquist and Suchard (2010)) to infer recombination. The number of inferred recombination events required to explain the history of the sample may be counted from the ARG. The Bayesian method SMARTIE (see Bloomquist and Suchard (2010)) can be applied to phylogenetic data to infer a single most probable ARG for a sample of sequences. However, current implementations of the ARG method in software such as BEAST may only be computed for a small number of sequences. In contrast, the NEST method could be applied to datasets of several hundred sequences, since the procedures used are essentially those employed for large-scale phylogeographic analyses in BEAST.

6.6.3 Future directions

Experimental studies have investigated the molecular basis of inter-subtype recombination in HIV-1. For example, Chin *et al*. (2005) compared *in vitro* rates of intra- and inter-subtype variation amongst subtype B and C viruses. Whilst recombination rates within subtypes B and C were found to be similar, the seven-fold lower rate of recombination observed between subtype B and C viruses was attributed to a three-nucleotide difference in the dimerization initiation signal (DIS) region between the two subtypes. The DIS region is involved in RNA packaging and it has been reported that HIV-1 group M subtypes split into two groups with respect to DIS sequence, with subtypes B and D having the motif GCGCGC and subtypes C, F, G, H and J possessing the motif GTGCAC (Sloth Andersen *et al.* 2003; Paillart *et al.* 2004). After creating subtype B and C vectors with matching DIS motifs, the level of inter-subtype B/C recombination increased four-fold. The observation that inter-subtype B/C recombination rates remained lower than the respective intra-subtype B and C recombination rates, even when the DIS motifs were matched, suggested that other minor restrictions to inter-subtype recombination exist. Having found that the frequency of template-switching was unaffected by whether or not the DIS motif was mismatched, Chin *et al*. (2005) suggested that coinfection with subtypes with mismatched DIS motifs resulted in the production of fewer heterozygous virions.

Based upon the results of the subtype B and C study, Chin *et al*. (2005) postulated that HIV-1 group M subtypes with different DIS motifs would exhibit lower inter-subtype recombination rates than subtypes which had the same DIS motif. In accordance with this hypothesis, experimental studies by Baird *et al*. (2006b) indicated that the frequency of recombination between subtype A and subtype D viruses, which belong to different DIS groups, was substantially lower than between two subtype A viruses. Given additional data, the NEST method could be used to test this hypothesis from phylogenies. For example, mixing between different subtypes could be assessed using counts of Markov jumps between particular pairs of subtypes, in a similar manner to the influenza NA subtype reassortment analysis of Chapter 5.

It may be noted that viruses from two individuals in the Kinshasa dataset were of p17 subtype D and gp41 subtype F, and one individual was infected with a virus of p17 subtype D and gp41 subtype G. This demonstrates the potential for inter-subtype recombination between subtypes from different groups *in vivo*. Furthermore, subtype B/C recombinant viruses (the subtypes considered in the inter-subtype recombination study of Chin *et al*. (2005)) have emerged in other natural populations, for example in East Asia (Piyasirisilp *et al.* 2000). In future, biological questions such as the level of multiple-infection which would be required in a population to have observed a given NEST rate (or recombination between particular subtypes) for a sample of sequences could also be investigated.

The study presented in this chapter focused on estimating the population-level rate of inter-subtype recombination across a phylogeny, based upon discordant p17 and gp41 subtypes. A novel method (NEST) which used discrete ancestral trait mapping on phylogenies to quantify the extent to which recombination has shaped the diversity of a sample of sequences was implemented. Given sequence data from several regions of the HIV genome, NEST could be used in future studies to compare the amount of inter-subtype recombination between different parts of the genome. The NEST method may also be used to quantify intra-subtype recombination in datasets of a single HIV-1 group M subtype, although intra-subtype recombination is more difficult to identify than inter-subtype recombination, due to a lack of distinguishing variation between parental and recombinant sequences (Salminen and Martin 2009). The method for defining clades from a certain cut-off point along the phylogeny (Section 6.4.4) could be implemented to estimate rates of intra-subtype recombination, although further work would be required to determine how the choice of cut-off affected the estimated rate of excess subtype transitions or the amount of recombination detected.

Chapter 7

Phylodynamics of viral epidemics

7 Phylodynamics of viral epidemics

7.1 Chapter Summary

Reconstructions of genealogies from viral sequence data contain information about the transmission dynamics of epidemics. In this section, a mathematical link is formulated between the coalescent theory of population genetics and standard epidemiological models. This allows epidemiological and demographic parameters to be inferred directly from viral sequence data, rather than from counts of infected individuals over time. It is shown that the coalescence rate for a viral phylogeny of sequences from different infected individuals is directly proportional to epidemic incidence, rather than the 'effective number of infections' assumed by previous researchers, of which it is a complex, non-linear function. The method also provides expressions for the expected distribution of phylogenetic cluster sizes under an SIR model of infection dynamics. In future the expected cluster size distribution may be used as a null distribution against which to detect departures from the assumptions of the SIR model, such as population structure or heterogeneity in transmission.

7.2 Chapter Aims

- Link viral sequence coalescence and epidemiological processes, so that SIR model parameters may be inferred from viral phylogenies (as presented by Volz *et al*. 2009 – see Chapter 11)
- Discuss how the method of Volz *et al*. (2009) compares to other methods for inferring epidemiological parameters from viral sequence data
- Explain how results are derived for the expected distribution of viral phylogenetic cluster sizes under SIR infection dynamics

7.3 Introduction

Phylogenetic trees describe evolutionary relationships between groups of molecular sequences. Lineages are represented by branches, whose lengths are a function of evolutionary time, and the nodes of the tree are the points at which lineages merge, backwards in time, towards a common ancestor. Divergence between lineages moving forwards in time results from the accumulation of genetic diversity due to nucleotide substitution. Coalescent theory (Kingman 1982) links the divergence times of lineages within a population (i.e. the shape of the genealogy) with the demographic history (i.e. the size over time) of that population. The coalescent effective population size for a Wright-Fisher population is defined as the value of N_e which provides the same distribution of coalescence times as would be obtained for the actual biological population under consideration (see Chapter 2, Section 2.12).

The original coalescent theory of Kingman (1982) assumed a constant population size. The requirement for the population size to be constant was relaxed to allow for deterministically varying population sizes by Griffiths and Tavaré (1994), using an integer-valued ancestor function $A(t)$ equal to the number of distinct ancestors of a sample at a time t in the past. Nee *et al*. (1995) also employed a lineage-counting approach in their use of the coalescent theory to infer past population dynamics. Later extensions of the approach of Nee *et al.* (1995) used the notion that the effective size of a population could change at coalescent events to introduce ‘skyline’ methods for visualizing changes in N_e as a function of time (Pybus *et al.* 2000; Strimmer and Pybus 2001) (see Chapter 2, Section 2.12).

Although the ‘standard’ coalescent theory results of Kingman assumed a selective neutrality, much work has been carried out on incorporating selection in the coalescent, for example the ‘structured coalescent’ (Kaplan *et al*. 1988; Hudson and Kaplan 1988), which was later extended by Nordborg (1999), and the ancestral selection graphs of Krone and Neuhauser (Krone and Neuhauser 1997; Neuhauser and Krone 1997). For a review of coalescent theory and selection, see Wakeley (2010). Coalescent-based analysis of pathogen sequence data which is likely to be under immune selection has been widely adopted, for example using the BEAST software of Drummond and Rambaut (2007).

If super-infection is rare (i.e. if individuals are typically only infected with one strain of a virus) and the viral mutation rate is sufficiently high relative to the rate of epidemic spread, then each lineage of a population-level (as opposed to within-host) viral phylogeny corresponds to a single infected individual with its own unique viral population. Genetic variability within the host may be addressed by sequencing multiple viral isolates and obtaining a consensus viral sequence for each sampled individual. Branching events in the tree (coalescence events when looking backwards in time) represent reproduction events in which both offspring have descendants in the sample. In the model, each coalescence event is closely associated with a transmission event (discussed in more detail by Volz (2012), although transmission is also occurring rapidly throughout the tree).

The relationship between transmission and phylogenetic branching for rapidly evolving infectious agents implies that the shape of a phylogeny, or the sequence of coalescence times, contains information about the transmission dynamics and epidemiological history of the pathogen (Holmes *et al.* 1995; Nee *et al.* 1995) (e.g. Figure 7.1). The study of the joint effect of evolutionary and epidemiological processes on shaping the diversity of genetic sequence data is known as phylodynamics (Grenfell *et al.* 2004). Measurably evolving RNA viruses, such as influenza and HIV, have been the subject of the majority of phylodynamic studies, owing to their rapid evolutionary rates (Duffy *et al.* 2008).

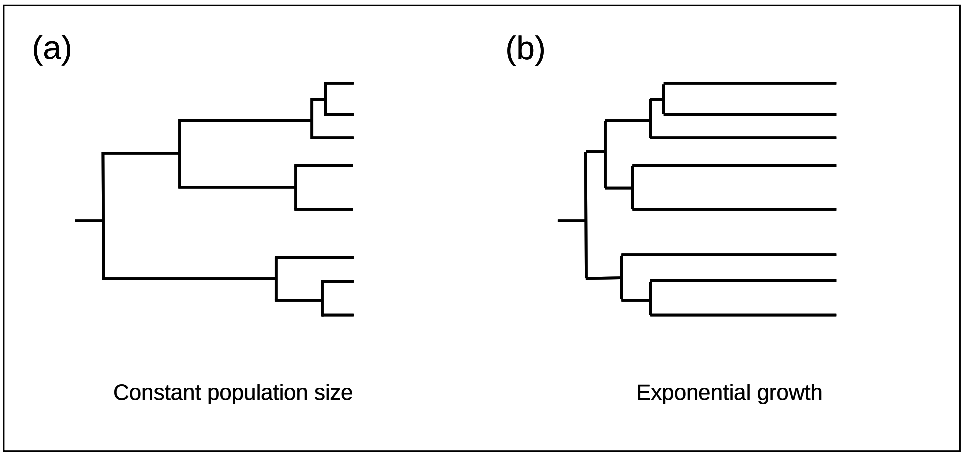


Figure 7.1
The shape of a viral phylogeny is affected by the underlying population dynamics. Trees are presented representing viral sequence data under two different models of population size: **(a)** constant population size and **(b)** an exponentially growing population. Coalescence times are pushed backwards under exponential population growth, compared to a constant population size. Based upon a figure by Grenfell *et al*. (2004).

Developments in phylodynamic inference are potentially of great public health importance, since estimating epidemiological parameters is crucial for informing strategies for disease prevention and control. However, many phylodynamic developments have been qualitative and formal links between the fields of pathogen evolution and transmission dynamics remain to be developed (Grenfell *et al.* 2004; Wilson *et al.* 2005; Pybus and Rambaut 2009). Many researchers have interpreted the coalescent effective population size from a viral phylogeny as the ‘effective number of infections’, a quantity assumed to be proportional to the size of the infected host population, without any formal links between these quantities having been derived (discussed by Frost and Volz (2010)).

In the first part of this chapter, a mathematical link between coalescent theory and epidemiological modelling is described, representing a step towards the integration of epidemiological and population genetic approaches. The method can be used to estimate the parameters of standard epidemiological models, such as SIR models, from the branching times for a phylogeny constructed from viral sequence data. The method provides a rapid and inexpensive method for estimating epidemiological parameters from genetic sequences taken from a small sample of an infected population early on in an epidemic, rather than relying on counts of the number of infected individuals over time. The relationship between coalescence rates, the rate of transmission (incidence) and the size of the infected population (prevalence) is also formalised. The second part of this chapter provides results for the expected distribution of phylogenetic cluster sizes.

7.4 Fitting epidemic models to viral sequence data

7.4.1 Introduction to epidemiological modeling

Standard epidemiological models, such as SIR, SI and SIS models, describe the infectious disease dynamics of a host population in terms of ordinary differential equations (ODEs). The entire host population of size N is divided into subpopulations, or compartments, into which individuals are classified according to their disease status. For example, in an SIR model the population is split into compartments of susceptible, infected and recovered individuals whose respective sizes at time t are denoted $N_S(t)$, $N_I(t)$ and $N_R(t)$. In this chapter, the situation with a constant overall population size is considered, i.e. $N_S(t) + N_I(t) + N_R(\mathrm{t}) = N$ for all t. Behaviour of the variables $S = N_S(t)/N$, $I = N_I(t)/N$ and $R = N_R(t)/N$ is modelled deterministically, forwards in time, in the limit $N \to \infty$, under the condition that S, I, $R >> 1/N$ (i.e. N_S, $N_{I,}$ $N_R >> 1$).

Under the classical mass-action SIR model (Kermack and McKendrick 1927; Bailey 1975), individuals move from the susceptible compartment to the infected compartment with rate $f_{SI}(t) = \beta S(t)I(t),$ and from being infected to recovered with rate $f_{IR}(t) = \gamma I(t)$, with all other rates of movement between compartments set to 0 (i.e. individuals cannot move from the recovered compartment to the infected or susceptible compartments, or from the infected compartment to the susceptible compartment). The transmission rate (f_{SI}) is proportional to the amount of contact between the susceptible and infected compartments, which is assumed to be determined by the product of their sizes. The SIR model may be written as:

$$\dot{S} = \frac{dS}{dt} = -\beta S(t)I(t)$$

$$\dot{I} = \frac{dI}{dt} = \beta S(t)I(t) - \gamma I(t)$$

$$\dot{R} = \frac{dR}{dt} = \gamma I(t)\,.$$

The ‘dot’ notation denotes the rate of change of a parameter, forwards in time, with respect to *t*.

The parameters of an SIR model can be estimated by fitting differential equations in the above form to counts of infected individuals (the ‘epidemic curve’) across a series of time-points (e.g. Wallinga and Teunis (2004a; 2004b), Lipsitch and Bergstrom (2004)). Of profound epidemiological significance is the quantity $R_0=\beta/\gamma$, known as the basic reproduction ratio, which is the expected number of infections caused by a single infected individual in a completely susceptible, immunologically naïve population (see Anderson and May (1991)). The value of R_0 determines whether an infectious disease outbreak has the potential to persist in a population: an outbreak cannot be sustained for $R_0 < 1$, whilst an epidemic may ensue when $R_0 > 1$. The estimated value of R_0 may be used in evaluating which control measures are necessary to prevent an epidemic, or to eradicate a particular infectious disease from a population.

Bringing R_0 to below 1 was the rationale for culling within the 3km exclusion zone around infected farms in the UK foot and mouth disease outbreak of 2001 (Woolhouse *et al.* 2001; Woolhouse and Donaldson 2001). In human populations, the point of interest is often to determine the proportion of a population to vaccinate, quarantine or administer prophylactic treatment to in order to curtail an epidemic. For example, Yang *et al*. (2009) simulated different vaccination strategies under a variety of epidemic scenarios in order to determine conditions for mitigating the

2009 H1N1 influenza pandemic. In practice, the ‘effective reproduction number’ R_e, is estimated rather than a strict R_0. R_e is the average number of secondary infections caused by an individual in a population which cannot be considered wholly susceptible due to previous exposure to the infectious agent, or past control measures (discussed by Amundsen *et al*. (2004) and Matthews and Woolhouse (2005)). Since R_e is directly proportional to S, strategies are often sought for reducing the size of the susceptible population in order to decrease R_0. Estimation of R_e can be carried out using various demographic or contact-tracing methods (e.g. Haydon *et al*. (1997)). For the early stages of an outbreak of a newly emerging infectious disease (to which there is no prior immunity and before large-scale prevention or control measures have been implemented) it is may be reasonable to assume that R_0 is being estimated. Note also that the SIR approach assumes no population structure and that mixing of individuals is homogeneous.

7.4.2 Coalescence and disease transmission

In this section, a formal link is developed between the coalescence rate of a sample of viral sequences and the rate of transmission of a virus under SIR dynamics. This link is used in Section 7.4.3 to derive an ordinary differential equation (ODE) to describe the coalescent process, which sheds light on the relationship between incidence, prevalence and the coalescence rate of a sample of viral sequences from an infected population. As outlined in Section 7.3, the coalescent theory of population genetics operates on a small sub-sample of a population of related individuals and models the merging of lineages backwards in time until a common ancestor has been reached. The merging of two lineages is known as a coalescence event, and coalescence is a stochastic process. In contrast, epidemiological compartment models such as SIR focus on the entire host population, operate forwards in time and, for the purposes of this chapter, are deterministic.

Consider a sample of size $n << N$*I (= N_{I}) genes taken from an infected population, of finite size N_{I}, of haploid individuals at time t. In a standard neutral Wright-Fisher model, when there are k lineages the ‘per-generation’ coalescence rate is proportional to $\binom{k}{2}$. (Note that the coalescence rate is $\binom{k}{2}.\frac{1}{N_I}$ per generation and some authors rescale time into units of N_I generations, so that the coalescence rate is $\binom{k}{2}$ per N_I generations.) See Nordborg (2000) or Wakeley (2009) for derivations and more detailed discussion.

In this chapter, the coalescence of lineages in a phylogeny constructed from viral sequence data will be considered under the assumptions that the sequences are consensus sequences from a small sample of individuals from the infected population, that super-infection is rare and that the mutation rate of the virus is fast relative to the transmission rate. The viral sequences will be assumed to have been sampled contemporaneously during an SIR epidemic. It is assumed that no recombination has taken place, so that the entire length of the genetic sequence has the same genealogical history and acts as a single locus. Each lineage may represent traversal of the virus over multiple infected hosts and transmission may not always result in a coalescence event being observed amongst the sample. This could be due to incomplete sampling of lineages, or because an individual has recovered, but failed to cause a subsequent infection before being sampled.

Given that a coalescence event occurs amongst the N_I infected individuals, the probability of observing the coalescence event amongst the sample of k observed lineages is given by the number of ways of choosing two lineages from the sample, divided by the number of ways of choosing two individuals from the entire infected population, i.e.:

$$\binom{k}{2} \Big/ \binom{N_I}{2} = \frac{k(k-1)}{N_I(N_I-1)}. \qquad (2)$$

By viewing coalescence and transmission as equivalent genealogical processes operating in different directions, the coalescence rate for a sample of sequences is given by the number of transmissions per unit time (f_{SI}), multiplied by the probability

(p_c) that a transmission results in a coalescent event being observed in the sample, i.e.:

$$\text{coalescence rate} = f_{SI} * p_c. \qquad (3)$$

The above relationship is used in Section 7.4.3 to derive an ODE to describe the coalescence process in terms of viral transmission parameters and the number of lineages as a function of time.

7.4.3 An ODE to describe the coalescent process

Consider a sample of consensus viral sequences sampled from individuals in an infected population at time T (with T fixed for now). Let the dimensionless variable $V(t)$ be the proportion of the entire population of size N at an earlier time t (with $t \leq T$) which has sampled progeny extant at time T. Since N does not change over time in the epidemic model described in Section 7.4.1 (whilst N_S, N_I and N_R are allowed to vary within the constraint that $N = N_S + N_I + N_R$), $V(t)$ is proportional to the number of ancestral lineages observed at time t. The quantity $V(t)*N$ is analogous to the integer-valued ancestor function $A(t)$ defined by Griffiths and Tavaré (1994) as the number of distinct ancestors of the sample at time t in the past. Note that the notation $V(t)$ is used in this chapter (where $A(t)$ is used for the same quantity in the manuscript of Volz *et al*. 2009) in order to distinguish between the quantity defined by Volz *et al*. (2009) and the ancestor function of Griffiths and Tavaré.

The function $V(t)$ (defined above) can be used to describe the coalescent process for a viral phylogeny as an ODE. Since $V(t)$ is proportional to the number of lineages in the phylogeny, the coalescence rate for the sample is given by the rate at which $V(t)$ changes with respect to t, moving towards the root of the tree, which can be linked with the transmission rate, as in equation (3). It is possible to use $V(t)$ to derive an expression for $p_c(t)$, the probability of observing a coalescent event in the genealogy for the sampled sequences, given that a transmission event has occurred in the population. This follows a similar argument to that outlined in Section 7.4.2, but here the expression is time-dependent since both the number of ancestors in the genealogy (proportional to *V*(*t*)) and the size of the infected population will vary over time. As the SIR model was defined in terms of a total population *N* in the limit as *N*→∞, the probability of observing a coalescent event in this limit is considered:

$$\begin{aligned} p_c(t) &= \lim_{N\to\infty} \binom{V(t)N}{2} \Big/ \binom{NI(t)}{2} \\ &= \lim_{N\to\infty} \frac{V(t)\big(V(t)N-1\big)}{I(t)\big(I(t)N-1\big)} \\ &= \lim_{N\to\infty} \frac{V(t)\big(V(t)-1/N\big)}{I(t)\big(I(t)-1/N\big)} \qquad (4) \\ &= \left(\frac{V(t)}{I(t)}\right)^2 . \end{aligned}$$

It must be noted that the above limit argument is an approximation, which requires that *V* >>1/*N*. The approximation will perform less well when the number of lineages is modest (i.e. towards the root of the tree).

From Section 7.4.2, the coalescence rate for the sample of sequences is equal to the rate at which individuals move from the susceptible to the infected population (the transmission rate, f_{SI}), multiplied by the probability (p_c) of observing a coalescent event in the sample, given that a transmission occurs. This is the rate of change of *V* moving forwards in time towards the present day, whereas moving backwards in time (towards the common ancestor of the sample) the rate of change of *V* is the negative of this quantity ($-f_{SI}*p_c$). The coalescence rate for the sample of sequence is analogous to the rate of change of *V*, backwards in time with respect to *t*, thus it is possible to write:

$$-\frac{dV}{dt} := \overset{-\bullet}{V} = -f_{SI}\,p_c = -f_{SI}\left(\frac{V(t)}{I(t)}\right)^2 . \qquad (5)$$

$V(t)$ can be found by integrating the above ODE backwards in time from T. Note that the ‘negative dot’ notation denotes the derivative of V moving backwards in time, which is simply the negative of the forward derivative.

As previously mentioned, the above approach is approximate, since V is a random variable which is treated as deterministic in the integration. Previous methods have been developed which allow the expectation of V (actually the ‘ancestor function’ VN) to be calculated as a function of time under a model of constant population size (Griffiths 1981; Tavaré 1984). When the population size is constant, Equation 5 appears to be a reasonable approximation for $VN > 1$ (Appendix E, Figure E1) and should behave well under the more stringent assumption of Volz *et al*. (2009) that $V >> 1/N$.

7.4.4 Finding maximum likelihood values of SIR parameters for known branching times

In Section 7.4.3, an ODE (Equation 5) was derived to describe the coalescent process in terms of the rate of change of the number of ancestral lineages in a viral phylogeny, which explicitly incorporated the transmission rate from an SIR model. In this section, a method for fitting an SIR model using the branching times inferred from a phylogenetic tree for the viral sequence data is presented. This allows estimates of epidemiological parameters to be obtained using maximum likelihood inference. In order to do this, the marginal distribution of coalescence times when there are n tips in the tree (i.e. n = number of sampled sequences) is considered. In addition to the assumptions made previously (i.e. that $V >> 1/N$ and that V could be treated as deterministic for integration), the coalescence times here are assumed to be drawn independently and identically distributed from the density function.

For a sample of n viral sequences, there will be a total of n-1branching (coalescence) events across the resulting phylogeny. Denote the branching times $t_1,\dots, t_{n-1}$, ordered so that t_1 is the time of the most recent branching event from the tips of the tree and

t_{n-1} is last branching event, i.e. the time at which the most recent common ancestor of the sample is reached. By definition, the value of V at the time of sampling (T) will be n/N for a sample of n individuals, and $V(t_{n-1}) = 1/N$. The quantity $V(T) - V(t_{n-1})$ is equal to $(n-1)/N$, i.e. proportional to the total number of coalescence events. Let the variable x denote an arbitrary point in time between t_{n-1} and T ($t_{n-1} \leq x \leq T$). Moving towards the root of the tree, the proportion of coalescent events that have taken place by time x is given by:

$$F(x) = \frac{V(T) - V(x)}{V(T) - V(t_{n-1})}. \qquad (6)$$

The function $F(x)$ takes values between 0 (when $x=T$) and 1 (when $x=t_{n-1}$) and is analogous to a cumulative distribution function for the distribution of coalescence times (Figure 7.2). Note that again the notation in this section differs slightly from that used by Volz *et al*. (2009). Simulations under a model of constant population size indicate that Equation 6 performs well as an approximation to the cumulative distribution function for coalescence times (Appendix E, Figure E2).

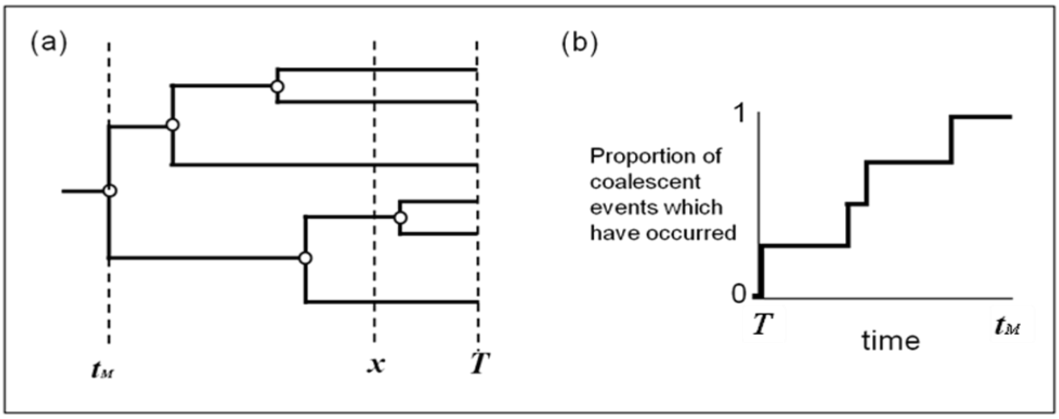


Figure 7.2
Proportion of coalescent events which have occurred across a tree. (a) Moving from the time of sampling (T) towards the root (at time t_M) of a tree constructed from n sequences, the number of coalescence events which have taken place changes incrementally from 0 (at time T) to n-1 (at time t_M). **(b)** Moving from the tips to the root of the tree, the number of coalescent events which have taken place as a proportion of the total number of coalescence events in the tree can be used to construct a density function for the distribution of coalescence times. This can be used to infer maximum likelihood values of epidemiological parameters based upon observed branching times. Note that the axes for figures (a) and (b) are in opposite temporal directions.

Differentiating *F*(*x*) with respect to *x* yields a probability density function for the coalescence times in the sample:

$$f(x) = -\dot{\bar{V}}/(V(T)-V(t_{n-1}))\,. \tag{7}$$

The density function *f*(*x*) can be used in calculating maximum likelihood estimates of epidemiological parameters, given inferred branching times $t_1,\ldots, t_{n\text{-}1}$ for a phylogeny. Assuming that every coalescence time is drawn independently and identically distributed from the density *f*, the probability of observing a particular set of branching times will be proportional to the product of the density evaluated at each branching time. Maximum likelihood estimates of SIR and demographic parameters θ may be obtained by maximising the following equation:

$$L(t_1,\ldots,t_{n-1}\,|\,\theta) = \prod_{i=1}^{n-1} -\dot{\bar{V}}(t_i)\,/[V(T)-V(t_{n-1})]\,. \tag{8}$$

In practice, the maximization may be performed by taking natural logarithms of both sides:

$$\ln[L(t_1,\ldots,t_{n-1}\,|\,\theta)] = \sum_{i=1}^{n-1} \ln[-\dot{\bar{V}}(t_i)\,/(V(T)-V(t_{n-1}))]\,. \tag{9}$$

7.4.5 Discussion of coalescence rate in relation to epidemiological parameters

As described in Section 7.3, previous researchers have noted that pathogen transmission is related to coalescence, or the number of lineages as a function of time. Despite this, phylodynamic studies have assumed that the coalescent effective population size (N_e) from a pathogen phylogeny is proportional to the disease prevalence, or ‘effective number of infections’, but have not explicitly investigated the relationship between coalescence rates and viral epidemiology. However, Equation 5 (Equation 2 of Volz *et al.* 2009) shows that the coalescence rate is directly proportional to the epidemic incidence (f_{SI}) and inversely proportional to the square of the prevalence (I^2). Frost and Volz (2010) thus argued that the generation of new lineages via transmission, rather than the number of infected individuals, largely determines the coalescence rate and has the major effect on shaping the phylogeny. It has been suggested that prevalence only affects the shape of the phylogeny indirectly, through sampling effects. For a fixed sample size, the level of sampling depends on the prevalence and sampling a higher proportion of the population causes a larger number of 'shallow' branching events to be observed (Volz *et al.* 2009; Frost and Volz 2010). The consequences of Equation 5 are that coalescence rates will be highest when the incidence is high and the prevalence is low, such as in the early stages of an epidemic.

Simulation and application to HIV sequence data has shown Equation 5 to accurately describe the dynamics of the number of lineages as a function of time. The method can also be used to estimate epidemic prevalence during the exponential growth stage (Volz *et al.* 2009). The method of Volz *et al.* (2009) has an advantage over skyline methods in that it does not rely on explicit estimates of the generation time τ, which is defined as the average length of time between an individual becoming infected and going on to infect another individual (Fine 2003). Generation time estimates may not be available, or reliable, for all pathogens. For a phylogenetic tree with branch lengths scaled in absolute time, the Bayesian skyline or skyride plot (see Chapter 2, Section 2.12) depicts the quantity $N_e\tau$ over time (Drummond *et al.* 2005). Although skyline methods may estimate the number of infectious individuals during exponential growth, when there is a linear relationship between transmission rate and prevalence ([3]), Frost and Volz (2010) assert that this relationship does not hold in general as τ changes over the course of an epidemic (for example, see Kenah *et al.* (2008)).

[3] During exponential growth, characterized by a population size at time t of $N = ae^{bt}$, the rate of change of N with respect to t (i.e. the rate of becoming infectious) is bae^{bt} ($=bN$).

Although Volz *et al*. (2009) found that one might distinguish SI from SIR dynamics under the framework described above, the method cannot be used to estimate R_0 without a prior estimate of the recovery rate (or average length of an infection). Other approaches which have estimated R_0 from the epidemic growth rate have also required independent estimates of an additional parameter: either the generation time (Wallinga and Lipsitch 2007; Grassly and Fraser 2008) or average duration of infectiousness (Pybus *et al.* 2001). For example, following the work of Pybus *et al*. (2001) many researchers have produced R_0 estimates for hepatitis C virus using the epidemic growth rate, *r*, and the average length of the infectious period, *D*, which are related by the equation $R_0 = 1 + rD$. Shortly after the 2009-H1N1 influenza outbreak was recognised, Fraser *et al*. (2009) obtained estimates of R_0 using estimates of the epidemic growth rate and the mean and variance of the human influenza generation time interval (Ferguson *et al.* 2005; Wallinga and Lipsitch 2007). Purely epidemiological methods also require estimates of the generation time in order to estimate R_0 from the epidemic curve in the absence of contact tracing data (Wallinga and Teunis (2004a; 2004b), Lipsitch and Bergstrom (2004)).

Non-coalescent methods have been developed for conducting epidemiological studies of viral sequence data, for example the birth-death approach of Stadler (Stadler 2009; Stadler 2010; Stadler *et al.* 2012). Birth-death models explicitly model the rate at which individuals become infectious (the birth rate) and the rate at which individuals become non-infectious, due to death, treatment or behavioral changes (the death rate). Transmission trees may be generated under this model. The sampling intensity ψ is explicitly accounted for in the model, as it is incorporated into the death rate. Since the birth and death rates can be estimated independently, R_0 can be estimated as the ratio of the birth rate to the death rate, directly from the viral sequence data, without a prior estimate of the recovery rate (Stadler *et al.* 2012). This is an advantage over the coalescent-based approaches such as that of Volz *et al*. (2009). A disadvantage of the birth-death approach compared to the coalescent is the assumption that the epidemic is undergoing exponential growth; birth-death models cannot fully describe dynamics across the course of an epidemic (Stadler *et al.* 2012), particularly towards the end of an epidemic in a finite population (Volz 2012). Implementation of methods such as those of Volz *et al*. (2009) and Stadler *et al*. (2012) in software such as BEAST will facilitate the comparison of coalescent and birth-death approaches to estimating epidemiological parameters from viral sequence data in different scenarios.

One scenario to which the method of Volz *et al*. (2009) could be applied is the early HIV-1 group M epidemic of West Central Africa, which was described in Chapter 6 (Section 6.3). A search of the literature did not find any previous estimates of R_0 for this region and period, obtained using either epidemiological or sequence data. Bayesian skyline plots for a sequence dataset sampled in 1984 in Kinshasa, the capital of the Democratic Republic of the Congo (DRC) (Appendix D, Figure D6), reflect the pattern observed for a larger collection of sequences sampled in the DRC at different time-points until 2005 (Worobey *et al.* 2008) (Appendix E, Figure E4). The skyline plots of Worobey *et al*. (2008) suggest a low but stable level of relative genetic diversity between the early 1900s and the mid-1950s, followed by a rapid increase in genetic diversity until the early 1980s, after which the genetic diversity started to plateau. It may be observed that there was an increase in the population size of the DRC (and of Kinshasa specifically) which was concomitant with the rise in the relative genetic diversity of HIV-1 group M between 1920 and 2000 (Appendix E, Figure E4). Compartment models may therefore need to account for changes in the overall human population size over time (e.g. Appendix E, Figure E5), when linking epidemiology with the coalescence of viral lineages in order to apply the methods of Volz et al. (2009). The birth-death model of Stadler *et al*. (2012), which allows birth and death rates, and thus R_0, to vary over time, may also be applied to the DRC HIV-1 sequence data. Temporal changes in R_0 could be visualized with a birth-death skyline plot (BEAST implementation in progress by Denise Kuhnert: http://code.google.com/p/bdssm-beast2/), and attempts could be made to explain such changes in terms of events such as urbanization and demographic expansion.

Whilst the methods of Volz *et al*. (2009) and Stadler *et al.* (2012) for estimating epidemiological parameters from viral sequence data have so far been applied predominantly to HIV, inference of parameters such as R_0 from avian influenza outbreak sequence data is yet to appear in the literature. Phylodynamic studies of avian influenza would require 'consensus' sequences for different infected farms (as were analysed in the H7N7 Netherlands outbreak by Bataille *et al*. (2011)) in order to calculate an inter-farm reproductive number (R_h). Estimates of epidemiological parameters from phylodynamic studies could be compared to estimates of inter-farm reproductive numbers obtained from epidemiological studies of avian influenza outbreaks, for example for the Italian H7N1 outbreak (Mannelli *et al.* 2007), or the Netherlands H7N7 outbreak (Boender *et al.* 2007).

Further to the results of Section 7.4, expressions for the coalescence rate over time could also be derived for SI and SIS models, as well as for more complex compartment models which portray infectious disease dynamics more realistically. For example, Volz *et al*. (2009) and Volz *et al*. (2012) (see Chapter 11) applied the method to a model which distinguishes between the transmission probabilities at early (acute) and later (chronic) stages of HIV transmission. The method presented by Volz *et al*. (2009) can also be extended to allow for heterochronous samples (i.e. samples taken from different at different points in time). This can be accomplished by performing a piece-wise integration of Equation 5 (Section 7.4.3) over intervals whose start- and end-points are given by consecutive ordered sample times.

7.5 Distribution of cluster sizes

In this section, expected properties of the distribution of phylogenetic cluster sizes under SIR dynamics are derived. Consider a phylogenetic tree constructed from consensus viral sequences isolated from *n* infected individuals. The *n* tips of the tree, which correspond to sampled individuals, can be grouped into clusters according to whether they are descended from a common lineage at some point *t* in the past. A '(t;T) cluster' is defined as the set of lineages at a later time *T* (in this chapter, the time of sampling) which are descended from a common ancestor at time *t*. The size of a (t;T) cluster is the number of progeny viral lineages at *T* of a lineage which exists at time *t* (Figure 7.3). Sequences belonging to the same phylogenetic cluster are often interpreted as representing epidemiologically linked individuals (e.g. Lewis *et al*. (2008) and Pilcher *et al*. (2008) for HIV). The properties of the distribution of cluster sizes are therefore of interest since they contain information about the transmission dynamics of an epidemic.

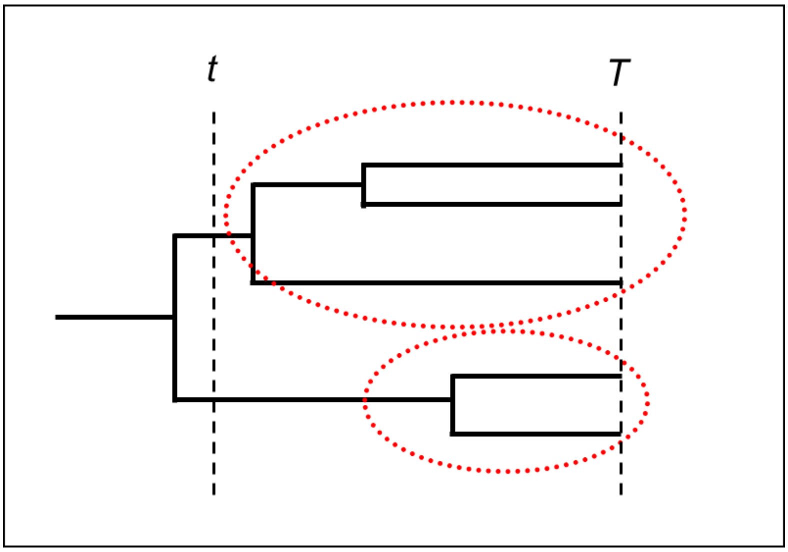


Figure 7.3
Definition and size of (*t*;*T*) clusters. There are a total of two (*t*;*T*) clusters (circled in red) for the values of *t* and *T* (indicated here by the dashed vertical lines). The top cluster is of size 3, whilst the bottom cluster has two extant lineages at time *T* and is therefore of size 2. The number and size of clusters depends upon the values of *t* and *T*.

7.5.1 Mean and variance of the distribution of cluster sizes

Consider a viral lineage at time *t*. Define $X_1(t;T)$ to be the number of progeny lineages of this lineage at the time of sampling, *T*, given that such progeny exist. Denote by $x_1(t;T)$ the expected size of a cluster at time *T*, for a lineage randomly selected at time *t*. Let $X_2(t;T)$ be a random variable describing the size of a cluster when clusters are selected with probability proportional to the cluster size, and denote $E(X_2)$ by x_2. Selecting clusters with probability proportional to their size yields the same distribution of cluster sizes as if infected individuals were sampled at time *T* and the size of the (*t*;*T*) clusters to which they belonged were determined.

Below, ODEs are derived which may be integrated backwards in time (from the tips of the tree towards the root) to find E(X_1) and E(X_2). Note that the assumptions and approximations made in Section 7.4 (i.e. V>>1/N and that the random variable V can be treated as deterministic for the integration) are also applied in this section, hence the expressions for the expected cluster sizes are themselves approximate. When compared to the exact results of Tavaré (1984), the behaviour of the approximation for x_1 (obtained using Equations 10 and 5 from this chapter) is seen to become less accurate towards the root of the tree, when the assumption that V>>1/N is violated (Appendix E, Figure E3).

The number of clusters of progeny viral lineages at time T, from lineages at an earlier time t, is given by $V(t;T)*N$, since the number of clusters will simply be equal to the number of lineages at time t which have progeny lineages extant at T. The number of lineages at T is $V(T;T)*N$. Note that, if all individuals in the population were sampled (as is the situation considered by Volz *et al*. (2009)) then $V(T;T) = I(T)$. The number of sampled infected individuals at time T is distributed across the $V(t;T)*N$ clusters, and the expected number of infected individuals per (t;T) cluster is obtained by dividing the number of infected individuals sampled at T by the number of clusters from time t with progeny lineages extant at T:

$$x_1(t;T) = \frac{V(T;T)N}{V(t;T)N} = \frac{V(T;T)}{V(t;T)}.$$

Rearranging gives:

$$V(t;T) = V(T;T)/x_1(t;T) \qquad (10).$$

Evaluating the derivative, backwards in time, with respect to t yields:

$$\overset{-\bullet}{V} = -\overset{-\bullet}{x}_1(t;T)\ V(T;T)/x_1(t;T)^2.$$

Using $\overset{-\bullet}{V} = -f_{SI}\left(\frac{V(t;T)}{I(t)}\right)^2$ from Section 7.4.3 combined with Equation (10) and rearranging, one obtains:

$$\overset{-\bullet}{x}_1(t;T) = f_{SI}\,V(T;T)/I(t)^2.$$

Now consider the size of a (t;T) cluster chosen with probability proportional to the size of the cluster, to which an individual at time T belongs. The size of a cluster changes when one cluster merges with another as one moves backwards in time towards the root of the tree. It is possible to derive an expression for the rate of change of x_2 (where x_2 is the expected size of a cluster selected with probability proportional to its size) as one progresses from the tips of the tree towards the root as follows: (i) two clusters merge at each coalescent event, and the average amount by which the cluster size increases upon merging with another cluster is given by x_1; (ii) since there are $V(t;T)*N$ clusters from time t, the probability that a given individual is involved in a coalescent event is proportional to $2/V$; (iii) from Section 7.4.2 and Section 7.4.3, the coalescence rate for the sample is $p_c * f_{SI}$. Multiplying the three quantities yields:

$$\overset{-\bullet}{x}_2 = -f_{SI}\,p_c\left({}^{2}\!/_{V}\right)\ x_1.$$

Using the relations $p_c = (V(t;T)/I(t))^2$ and $V(t;T) = I(T)/x_1(t;T)$ and rearranging gives:

$$\overset{-\bullet}{x}_2 = 2\,\overset{-\bullet}{x}_1\,.$$

The expressions for $\overset{-\bullet}{x}_1$ and $\overset{-\bullet}{x}_2$ can both be solved by integrating in reverse time with the initial conditions $x_1(T;T) = 1$ and $x_2(T;T) = 1$ respectively (as, by definition, each (T;T) cluster of individuals will always consist only of one lineage).

It is also possible to derive an expression for the variance of X_1, by noting that

$$E(X_1^2) = \sum_i i^2 P\{X_1 = i\} = \left(\sum_i iP\{X_1 = i\}\right)\left(\frac{\sum_i i^2 P\{X_1 = i\}}{\sum_i iP\{X_1 = i\}}\right) = x_1.\left(\frac{\sum_i i^2 P\{X_1 = i\}}{\sum_i iP\{X_1 = i\}}\right).$$

X_2 is the size of a cluster whose probability of selection is proportional to its size, hence one can write:

$$P\{X_2 = i\} = \frac{iP\{X_1 = i\}}{\sum_j jP\{X_1 = j\}}$$

and

$$x_2 = E(X_2) = \frac{\sum_i i^2 P\{X_1 = i\}}{\sum_j jP\{X_1 = j\}}.$$

Thus, $E(X_1{}^2) = x_1 x_2$ and $Var(X_1) = E(X_1^2) - [E(X_1)]^2 = x_1 x_2 - x_1^2$.

7.5.2 Finding the nth moment of the cluster size distribution

Let $X_1(t;T)$ be the random variable denoting the number of progeny at time T of a randomly chosen lineage at an earlier time t. The first three moments of the distribution of cluster sizes are given as follows:

0th moment: $M_0 = E(X_1{}^0) = E(1) = 1$
1st moment: $M_1 = \mathrm{E}(X_1) = x_1$ by definition
2nd moment: $M_2 = E(X_1{}^2) = x_1 x_2$ (derived in Section 7.5.1).

In general, the n^{th} moment, M_{n}, of a random variable X is given by:

$$M_n = E(X^n) = \sum_i i^n P\{X = i\}.$$

In this section, a general expression is given which, when integrated backwards in time with appropriate initial conditions, means that it is possible to find any moment of the cluster size distribution. This expression is derived by considering the expected change in the sizes of clusters as they merge towards the root of the tree. It should be noted that the approximations and assumptions described previously in this chapter are also made in this section, hence the expression for the moments of the cluster size distribution is an approximation. However, lineages are exchangeable under the neutral coalescent, thus the distribution of coalescence times is independent of the topology. Therefore, a random topology may be generated, with coalescence times sampled afterwards from an exponential distribution for waiting times in order to obtain branch lengths. This property could be exploited in the future to obtain an exact cluster size distribution, using results from the population genetics literature such as those of Tavaré (1984), for the probability of transition from k to K lineages between times t and T, and Donnelly (1986), for describing genealogical processes moving forwards in time from the root of the tree.

Consider a cluster of size i. Upon merging with a cluster of size j, the n^{th} power of the cluster size will increase from i^n to $(i+j)^n$. An expression is derived for the expected change in the n^{th} moment when a cluster of size i merges with a cluster of size j:

*n*th moment for cluster of size (*i*+*j*) – *n*th moment for cluster of size *i*

$$
\begin{aligned}
&= \sum_i \sum_j P(X_1 = i)P(X_1 = j)(i+j)^n - \sum_i P(X_1 = i)(i^n) \\
&= \sum_i \sum_j P(X_1 = i)P(X_1 = j)(i+j)^n - M_n \\
&= \sum_i \sum_j P(X_1 = i)P(X_1 = j)\sum_{m=0}^{n}\binom{n}{m} i^{n-m} j^m - M_n \, * \\
&= \sum_i \sum_{m=0}^{n} P(X_1 = i)\binom{n}{m} i^{n-m} \sum_j P(X_1 = j)(j^m) - M_n \\
&= \sum_i \sum_{m=0}^{n} P(X_1 = i)\binom{n}{m} i^{n-m} M_m - M_n \\
&= \sum_{m=0}^{n}\binom{n}{m} M_m \sum_i P(X_1 = i)(i^{n-m}) - M_n \\
&= \sum_{m=0}^{n}\binom{n}{m} M_m M_{n-m} - M_n \\
&= \sum_{m=0}^{n-1}\binom{n}{m} M_m M_{n-m}
\end{aligned}
$$

* Using the binomial expansion.

The *n*th moment of the cluster size distribution can be found by integrating the equation for the rate of change of M_n, which is the product of the coalescence rate $f_{\mathrm{SI}}(V/I)^2$, the factor 1/*V* ([4]) and the expected change in the *n*th moment when two clusters merge:

$$\dot{\bar{M}}_n = f_{SI}\frac{V}{I^2}\sum_{i=0}^{n-1}\binom{n}{i} M_i M_{n-i}\ .$$

[4] The factor 1/*V* accounts for the probability that a given lineage takes part in a coalescence event (probability = 2/*V* from above), but is not the lineage that is lost at the coalescence event (probability = 1/2).

In performing the integration to find higher moments using this expression, the initial condition $M_n(T) = 1$ may be used, which uses the property that at the time of sampling all clusters will be of size 1.

It may be noted that the problem addressed by Volz *et al*. (2009) could be interpreted as a more general coalescent scenario, rather than merely one of infectious disease dynamics, and may be compared to more traditional population genetics studies. For example, Barton and Etheridge (2004), considered the effect of selection on genealogies for a linked, neutral locus and derived a generating function for pairwise coalescence times. Whereas the population considered by Volz *et al*. (2009) (the number of individuals in the infected population) varied deterministically according to the SIR model, under the approach of Barton and Etheridge (2004) the population (reported in terms of allele frequencies rather than the number of infected individuals) varied stochastically according to a diffusion process. However, the analytical complexity of the approach of Barton and Etheridge prohibited them from considering the distribution of cluster sizes or higher order moments. In contrast, the recursive formula for the rate of change of an arbitrary moment of the cluster size distribution presented by Volz *et al*. (2009) is simple to solve computationally. In the future, the work of Volz *et al*. (2009) could help to shed light on questions from outside the field of epidemiology, for example general colonisation problems.

A potential application for the expressions for the moments of the cluster size distribution described in Section 7.5.1 and 7.5.2 would be in generating null distributions for phylogenetic cluster sizes under an SIR or other epidemic model. Developments could enable formal statistical testing for departure from the assumptions of the epidemiological model. For example, the actual distribution of cluster sizes could be compared with the null distribution in order to detect marked transmission heterogeneity in terms of variance in the number of progeny lineages.

Chapter 8

Thesis summary and discussion

8 Thesis summary and discussion

Avian influenza viruses present an epidemiological and economic threat on a global scale, being associated with substantial losses to the poultry industry, as well as having the potential for transmission to swine and humans. However, many questions remain unanswered as to how avian influenza viruses evolve in their natural reservoir of wild aquatic fowl, and in non-natural avian hosts such as terrestrial poultry (Nelson and Holmes 2007). Analysis of avian influenza RNA sequences, using phylogenetic methods in conjunction with high-performance computing, can help to shed light on important evolutionary questions and enhance our understanding of the pathogen and its behaviour. The development of new techniques for analysing influenza virus evolution is also an important area for research, allowing a larger number of questions to be addressed and more complex hypotheses to be tested. Novel methods may also find applications in the analysis of other rapidly evolving viruses associated with a large public-health burden, such as HIV.

H7 and H5 are the only haemagglutinin (HA) subtypes which have been found to be highly pathogenic in chickens. H7 viruses present an immediate threat to poultry, and potentially to humans if they are able to cause infection and become easily transmissible from person-to-person. In this thesis, I focused on the analysis of avian influenza sequences from viruses of the H7 HA subtype, which has been under-studied compared to H5. In particular, I investigated the evolution of character traits along phylogenies. I used an extension of the stochastic mutational mapping methods of Nielsen (Nielsen 2001; Nielsen 2002) to calculate the ratio of the rate of non-synonymous substitution to the rate of synonymous substitution (d_N/d_S) along different parts of a phylogeny associated with different character traits, such as viral NA subtype or avian host type. I then investigated the use of Bayesian phylogeographic methods in the BEAST software (Lemey *et al.* 2009) for mapping discrete ancestral character traits, with the aim of quantifying reassortment across phylogenies, in particular the rate at which the HA of H7 avian influenza viruses is exposed to different NA subtypes. Discrete trait mapping methods and transition-counting methods (Minin and Suchard 2008a; Minin and Suchard 2008b; O'Brien *et al*. 2009; Talbi *et al*. 2009) were then employed for quantifying inter-subtype recombination between opposite ends of the HIV-1 genome. I also provided a detailed technical description of methods for estimating epidemiological parameters from genetic sequence data. I discussed the applicability of these, and other methods with a similar aim, to avian and human influenza and other pathogens. In the following paragraphs I will summarise the studies described in this thesis, and the main conclusions that I have drawn from them. I will discuss how such findings reconcile with current knowledge of avian influenza, outline questions which remain unresolved and identify how the studies undertaken in this PhD could lead to future investigations.

8.1 Thesis summary

Chapter 1 of this thesis was an introduction to avian influenza viruses and their evolution. In Chapter 2, I gave a detailed description of methods which can be applied to investigate the evolution of influenza and other viral pathogens, many of which were employed in the studies described in this thesis.

In Chapter 3, I undertook a molecular evolutionary analysis of all available avian influenza H7 HA sequences in the NCBI database: a total of 470 sequences sampled between 1927 and 2009. At the time that this analysis was performed, it was the most comprehensive dataset available in terms of number of sequences and span of sampling times, with the preceding study of avian H7 HA by Banks *et al*. (2000) having examined just 54 sequences which were all sampled in 1999 or earlier. Phylogenetic analysis of the avian H7 HA sequences revealed two main clades: the 'Eurasian' clade (sequences from Eurasia, Africa, Australia and New Zealand) and the 'American' clade (North and South American sequences). The lack of exchange of H7 HA between the Eurasian and American clades had been identified in the earlier dataset of Banks *et al*. (2000) and the results of this chapter indicate that the pattern has persisted over time, and despite greater depth of sampling. I also observed clustering into smaller geographical regions, for example the existence of a distinct Australian H7 HA clade within the Eurasian clade, and of distinct North and South American clades within the American clade, which has recently been confirmed for H7 HA by other researchers (Bulach *et al.* 2010; Lebarbenchon and Stallknecht 2011). In addition, I provided strong phylogenetic evidence that the transport of caged birds might facilitate the global spread of avian influenza, identifying two independent instances in the 1990s where European-like viruses had been sampled from imported birds in North America. These findings lend weight to the hypothesis that the transport of caged birds might be the most likely route for highly pathogenic H5N1 to enter America from Eurasia in the future (Webster *et al.* 2007). Further evidence of the geographical isolation between avian influenza viruses in the Americas and Eurasia, Africa and Australasia was provided by the finding of distinct sets of HA cleavage site motifs for the two major clades.

In Chapter 4, I investigated the selective pressure experienced by avian influenza H7 HA using an extension of the Bayesian stochastic mutational mapping methods of Nielsen (2001, 2002). Rather than merely counting numbers of synonymous and non-synonymous changes along branches, the method used a rescaling to account for the degeneracy of the genetic code and the time spent in each codon state along the tree, enabling the ratio of the rates of non-synonymous and synonymous substitution (d_N/d_S) to be calculated. In Chapter 3, I had shown that different background NA viral subtypes were distributed across the H7 HA phylogeny – a signature of reassortment – rather than forming distinct clades according to the NA subtype. I also showed that sequences from wild and domestic hosts were intermixed across the phylogeny and, consistent with previous researchers, I observed that highly pathogenic (HP) sequences were distributed across the phylogeny rather than forming a separate HP lineage (Rohm *et al.* 1995; Banks *et al.* 2000). By using parsimony mapping to infer ancestral viral NA subtypes on the H7 HA trees, I was able to compare d_N/d_S along parts of the phylogeny corresponding to lineages with different viral NA subtypes and investigate whether the selective pressure experienced by avian influenza H7 HA varied between different NA subtype backgrounds.

The results of the study of H7 HA selective pressure indicated that the average d_N/d_S across the HA1 functional region was higher on the N2 NA background (i.e. in H7N2 viruses) than on N1, N3 and N7 NA backgrounds. No substantial difference in d_N/d_S was detected between HP and LP viruses, or between viruses from terrestrial or aquatic birds. These results would be consistent with the hypothesis that genetic interactions between HA and NA can lead to a difference in the selective pressure experienced by H7 HA on different NA subtype backgrounds, possibly due to an experimentally confirmed requirement to evolve to achieve a functional balance between the activity levels of the proteins they encode. However, the heterogeneity of the H7 HA dataset composition in terms of the demographic scenario from which the sequences were obtained means that it is difficult to attribute the higher d_N/d_S for H7N2 HA simply to the viral NA subtype background. In particular, the majority of the H7N2 sequences were from the live bird markets, in which the nature of selection might differ from the settings in which sequences from the other NA background subtypes were sampled (for example where more sequences might be from wild birds).

In Chapter 5, I further investigated reassortment between HA and NA in avian influenza H7 viruses, using the discrete trait mapping in BEAST. I used the methods of Lemey *et al*. (2009) to map the NA subtype of the virus onto the H7 HA phylogeny and identified significantly non-zero rates of transition from one NA subtype to another using Bayesian stochastic search variable selection (BSSVS). I also used ‘Markov jumps’ methods for counting labelled transitions (Minin and Suchard 2008a; Minin and Suchard 2008b; O'Brien *et al.* 2009; Talbi *et al.* 2009) to quantify discrete trait transition across the phylogeny samples. I showed that reassortment events producing viruses with different NA subtype backgrounds occurred more frequently in H7 avian influenza than switching between wild and domestic avian hosts, or between low and high viral pathogenicity. Although the majority of host-switching events were from wild to domestic birds, the analyses provided quantitative evidence for spill-over back from poultry into wild birds, such as has been previously suggested for H5N1 (Chen *et al.* 2005; Feare 2010). In contrast, transition between viruses of low and high pathogenicity was essentially unidirectional from LP to HP. I also concluded that transition-counting might be the most appropriate method for quantifying processes such as reassortment across phylogenies, and on that basis used Markov jumps counting to investigate recombination in HIV-1 group M in Chapter 6.

In Chapter 6, I used the discrete trait mapping methods of Lemey *et al*. (2009) with transition-counting methods (Minin and Suchard 2008a; Minin and Suchard 2008b; O'Brien *et al.* 2009; Talbi *et al.* 2009) to quantify inter-subtype recombination across HIV-1 group M phylogenies. The method was applied to a previously published dataset from Kinshasa, the capital of the Democratic Republic of the Congo, in West Central Africa, which is believed to be the epicentre of the HIV-1 epidemic. Amongst a sample of *gag* p17 and *env* gp41 sequences from 57 patients, I found that on average 10.55-12.18 extra ancestral subtype transitions were required to map patients' viral subtypes from one end of the genome onto the tree for the opposite end of the genome, compared to onto the tree for the correct end of the genome. Such events reflect the amount of inter-subtype recombination which has occurred in the population to shape the phylogenies for different regions of the genome. By scaling the Number of Excess Subtype Transitions (NEST) to account for the time across the phylogeny, I obtained an estimate of the rate at which such events could be detected by comparing phylogenies for the p17 and gp41 regions, of 7×10^{-3} per lineage, per year. NEST is a more sophisticated statistic for measuring inter-subtype recombination in a sample of sequences than simply calculating the proportion of individuals with discordant subtypes in different regions of the genome, of which many instances could represent infection with already recombinant viruses rather than independent inter-subtype recombination events. I also discussed potential developments of the NEST method, including the possibility of reconciling it with population-genetic measures of recombination rates, and proposed further investigations which could be carried out using variants of the method.

In Chapter 7, I provided a detailed technical description of a method for linking epidemiological models with viral phylogenies under the coalescent theory. The method allows viral transmission rates to be estimated directly from the branching times of a viral phylogeny for a small sample of infected individuals in a population. The advantage over traditional methods for fitting epidemiological models is that this method allows epidemiological parameters to be inferred rapidly and at relatively little logistical expense, compared to fitting an epidemiological model to counts of infections over time. I discussed the relative merits of the approach of Volz *et al*. (2009) and other recent methods for inferring epidemiological parameters from sequence data for different pathogens, and described challenges which remain to be addressed. I also made specific reference to the use of such methods for elucidating the early epidemiological behaviour of HIV-1 group M in West Central Africa, using the dataset studied in Chapter 6.

8.2 Future directions

As well as performing traditional molecular evolutionary analysis of genetic sequence data, in this thesis I have used novel methods for analysing viral evolution and contributed to their development. For example, stochastic mutational mapping methods were extended to enable d_N/d_S to be estimated across different parts of a phylogeny in Chapter 4, and discrete phylogeography methods were applied to investigate reassortment and recombination in Chapters 5 and 6. These developments may facilitate future exploration of the evolutionary dynamics of avian influenza and other rapidly evolving RNA viruses. In the following section, I will discuss advances in statistical phylogenetics, computation and data availability which have taken place during the course of this PhD. I will consider how the work presented in this thesis fits in with such developments and how the conclusions and limitations of the thesis highlight important areas for future work.

Bayesian phylogenetics has progressed rapidly since this PhD commenced. Many additional analysis options have been implemented in the BEAST software; for example, simulation-free methods for calculating d_N/d_S from labelled transitions along phylogenies (O'Brien *et al.* 2009) (see Section 2.10.3) could now provide an alternative, and more computationally efficient, method to the stochastic mutational mapping employed in Chapter 4. Alongside such methodological advances have come computational developments such as the use of graphics processing units (GPUs), whose highly parallel architecture, consisting of many processing cores, has been exploited to substantially accelerate statistical phylogenetic inference. The high-performance BEAGLE library can efficiently perform the likelihood calculations that underpin much of statistical phylogenetics by making effective use of available computer hardware, including GPUs (Ayres *et al.* 2012). The BEAGLE library was employed in the discrete trait mapping analyses performed towards the end of this PhD, and in the future the use of many-core algorithms could facilitate the routine use of more complex and biologically realistic models. For example, GPUs could allow full-codon models to be implemented (Suchard and Rambaut 2009), which could be used to investigate selective pressure (Yang *et al.* 2000) as an alternative to nucleotide mapping carried out in Chapter 4.

All avian influenza virus sequence data analysed in this thesis were obtained from the publically accessible NCBI database. Whilst the NCBI database is a rich resource for influenza sequence data, it was noticed that many sequences, particularly the earlier sequences, were inconsistent in the way in which they were labelled. This resulted in a loss of information when sequences had to be excluded from analyses. For example, even after exhaustive literature-searching, which is time-consuming in itself, the taxonomic order and status (wild or domestic) of the avian host could not be ascertained for almost 10% of the avian H7 HA sequences examined in Chapter 3. There is a risk of bias in studies which combine data from different sources, if all of the sequences submitted by a particular researcher or study must be omitted from an analysis due to inconsistent or inadequate labelling.

One of the most important recent advances in Bayesian phylogenetics has been the combined analysis of genetic sequences and associated discrete (or continuous) trait data. Whilst discrete trait mapping was a major focus of this thesis, phylogeographic questions were not addressed; a barrier to such studies is the lack of systematically sampled avian influenza sequence data from different locations, which could lead to substantial biases (as suggested in Chapter 5, Section 5.5.2, in using Bayesian stochastic search variable selection to test whether rates of NA subtype transition were significantly non-zero). With sufficient spatiotemporal coverage, phylogeographic methods such as those of Lemey *et al*. (2009, 2010) may be used to investigate disease transmission on different geographical scales (e.g. Raghwani *et al*. 2011). Furthermore, Bayes factor testing between models using different spatially-informed prior distributions, incorporating distances and/or rates of movement between locations, and their population densities, can elucidate the role of migration in the spread of a disease (Gray *et al.* 2011; Allicock *et al.* 2012). In future, the role of bird migration in the spread of avian influenza could be investigated by performing phylogeographic analyses with a spatial diffusion prior informed by known migratory flyways (see Chapter 1, Figure 1.3), perhaps using information from satellite tracking of wild birds (e.g. Gilbert *et al*. (2010)). For example the relative contribution of longitudinal and latitudinal bird movement to influenza spread in North and South America could be investigated. In Chapters 5 and 6, discrete trait mapping methods were employed to study reassortment in influenza and inter-subtype recombination in HIV, demonstrating how techniques which have traditionally been used for phylogeographic analysis may be adapted to address novel evolutionary questions. Other potential uses for discrete trait mapping would be to investigate the evolution of phenotypes such as antiviral drug resistance (in birds or humans) along phylogenies. Traits such as drug resistance and adaptation of avian influenza viruses to domestic hosts could also be addressed by using Bayesian graphical models (e.g. Lycett *et al*. 2009 – see Chapter 11).

Although the interface between evolutionary biology and epidemiology, discussed in Chapter 7, is a rapidly progressing field, little has been done to apply this theory to avian influenza outbreaks. For example, during and after the Italian H7N1 outbreak, sequence data (Banks *et al.* 2001) and epidemiological data (Mannelli *et al.* 2006; Mannelli *et al.* 2007) were reported separately by different researchers. Since there was no link between the epidemiological and evolutionary data, and no information about the locations from which sequences were sampled, this prohibited a phylogeographic or combined evolutionary and epidemiological analysis. Even if the conventions for influenza virus nomenclature (Chapter 1, Section 1.3) are adhered to, properties such viral pathogenicity, whether the avian host was wild or domestic, or the precise geographical sampling location, are not routinely reported in the sequence databases. In order to facilitate large-scale analyses of avian influenza which combine genetic sequences and other information from numerous studies to test complex evolutionary and epidemiological hypotheses, it would be beneficial to have a protocol in place so that additional information was made available in the sequence database at the time of submission, and could be added to at a later date, for example following laboratory testing for pathogenicity.

Recent studies have used the genetic relationships between sequences to infer inter-farm transmission networks for avian influenza (Bataille *et al.* 2011) and foot-and-mouth virus (Cottam *et al.* 2008). Future work could involve the incorporation of poultry network information (e.g. Nickbakhsh *et al*. (2011)) into spatiotemporal genetic sequence analyses to examine the spread of avian influenza within a country or region. However, the ability to perform such analyses and draw sound conclusions from them ultimately rests upon the quality of data in terms of the depth of sample coverage in space and time. Previously, sufficient sequence data were not available in the publically available avian influenza databases to perform detailed evolutionary-epidemiological analyses on specific avian influenza outbreaks. However, some data from the 2003 H7N7 avian influenza outbreak in the Netherlands (Bataille *et al.* 2011; Jonges *et al.* 2011; Ypma *et al.* 2012) has recently been added to the GISAID database (http://gisaid.org). Given the high level of coverage (sequences from 70-90% of infected farms), the Netherlands avian influenza outbreak data would be an excellent setting for phylogenetic network studies and for investigating the use of methods for inferring epidemiological parameters from genetic sequence data, for example using the methods described in Chapter 7.

At the same time as advances in statistical phylogenetic methods and computation have taken place, the amount of influenza sequence data available for analysis has increased dramatically. The overall number of influenza sequences available from the NCBI Influenza Virus Resource has quadrupled since the start of this PhD (NCBI 2012). In the past, often only the HA has been sequenced, and HA evolution has thus been the focus of this thesis and other studies of avian influenza evolution (e.g. Banks *et al*. (2000), Lebarbenchon and Stallknecht (2011)). The advent of next-generation sequencing technologies means that today many more full-genomes are being sequenced. Analysis of full-genome influenza sequence data has traditionally been complicated by the presence of reassortment, which means that different phylogenetic trees may be required for different segments and hence that the segments must be analysed separately (e.g. Vijaykrishna *et al*. (2008a)). A class of models which could prove useful for analysing the growing amount of full-genome influenza sequence data are hierarchical phylogenetic models (HPMs), which have recently been implemented in BEAST. HPMs may be used to analyse sequence data which has been partitioned into segments, and the precision of parameter estimates can be increased by pooling information across the partitions without requiring congruence between the phylogenetic trees for different segments (Suchard *et al.* 2003). Furthermore, the use of hierarchical prior models could allow influenza segments to be grouped into specific rate classes (Bloomquist and Suchard 2010), which could help to address the requirement to incorporate genetic interactions between segments identified in Chapter 4.

Finally, through examining the distribution of avian influenza sequences in the NCBI database, I have identified geographical regions which are chronically under-sampled in both wild and domestic avian hosts: in particular South America, Central America and Africa. Amongst the full-length avian influenza H7 HA dataset of over 400 sequences described in Chapter 3, just 9 sequences were available from South America, along with 2 sequences from Central America and 5 from Africa. The wild bird influenza dataset described in Section 5.6 contained over 1,400 polymerase (PB2) sequences from North America, whereas only 11 sequences were available from Central America and South America. Whilst the North and South American sequences formed distinct clades in the analysis of avian influenza H7 HA in Chapter 3, the nine available South American sequences clustered with the North American sequences in three different clades across the wild bird PB2 phylogeny (not shown). Mixing of viruses between wild birds in North and South America would not be surprising given the potential for North-South spread of the virus along the Americas flyway (Chapter 1, Figure 1.3).

The extent to which North and South American avian influenza sequences cluster phylogenetically, and the role of migration in the spread of avian influenza, cannot be formally assessed without additional sequence data from Central and South America. Furthermore, the under-sampling and lack of sequencing of influenza viruses from both wild and domestic birds in Central and South America means that these are potential regions in which surveillance could reveal previously unobserved influenza virus diversity. The recent discovery of a distinct influenza lineage (subtype H17 HA) amongst bats in Guatemala (Tong *et al.* 2012) highlights the importance of surveillance in wild mammalian populations, as well as in domestic mammals, as was demonstrated by the long period of un-sampled influenza diversity in swine which preceded the 2009 H1N1 pandemic (Smith *et al.* 2009).

8.3 Concluding remarks

In this thesis, I addressed issues such as epistasis and genetic interactions between influenza RNA segments, the ability to quantify reassortment or recombination from phylogenies and the inference of evolutionary parameters from genetic sequence data. My results indicate that avian influenza H7 HA remains genetically distinct between the Americas and Eurasia, consistent with known migratory flyways, with further subdivision at the inter-continental level. On smaller scales, reassortment with different NA subtype backgrounds, switching between wild and domestic avian hosts and the emergence of highly pathogenic viruses from low pathogenic forms may be observed over the same timescale as the accumulation of nucleotide substitutions, with NA-subtype reassortment occurring most frequently. At the RNA segment level, I obtained results consistent with the idea that evolution may be affected by the genetic background in which a segment finds itself. In this chapter, I have discussed my findings in the context of recent developments in our understanding of, and ability to study, viral sequence evolution. I outlined how the methods from this thesis could be applied to further studies of viral evolution and highlighted how particular studies are prohibited by a lack of appropriate data or systematic sampling. It is hoped that, in the future, the results and ideas presented in this thesis may be used to further our knowledge of infectious diseases which present a significant threat to human and animal health.

Chapter 9

Bibliography

Abbas, M. A., E. Spackman, D. E. Swayne, Z. Ahmed, L. Sarmento, *et al.* (2010). Sequence and phylogenetic analysis of H7N3 avian influenza viruses isolated from poultry in Pakistan 1995-2004. *Virology Journal* **7**.

Abecasis, A. B., P. Lemey, N. Vidal, T. de Oliveira, M. Peeters, *et al.* (2007). Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: Subtype G is a circulating recombinant form. *Journal of Virology* **81**(16): 8543-8551.

Air, G. M. (1981). Sequence Relationships among the Hemagglutinin Genes of 12 Subtypes of Influenza-a Virus. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* **78**(12): 7639-7643.

Air, G. M., M. C. Els, L. E. Brown, W. G. Laver and R. G. Webster (1985). Location of Antigenic Sites on the 3-Dimensional Structure of the Influenza N2 Virus Neuraminidase. *Virology* **145**(2): 237-248.

Akaike, H. (1974). New Look at Statistical-Model Identification. *Ieee Transactions on Automatic Control* **Ac19**(6): 716-723.

Akaike, H. (1983). Information measures and model selection. *International Statistical Institute* **44**: 139-149.

Akey, B. L. (2003). Low-pathogenicity H7N2 avian influenza ity outbreak in Virginia during 2002. *Avian Diseases* **47**: 1099-1103.

Alexander, D. J. (2000). Highly pathogenic avian influenza. *OIE Manual of Standards for Diagnostic Tests and Vaccines*. W. O. f. A. H. OIE. Paris**:** 212–220.

Alexander, D. J. (2003). Report on avian influenza in the Eastern Hemisphere during 1997-2002. *Avian Diseases* **47**: 792-797.

Allicock, O. M., P. Lemey, A. J. Tatem, O. G. Pybus, S. N. Bennett, *et al.* (2012). Phylogeography and Population Dynamics of Dengue Viruses in the Americas. *Molecular Biology and Evolution*.

Amundsen, E. J., H. Stigum, J. A. Rottingen and O. O. Aalen (2004). Definition and estimation of an actual reproduction number describing past infectious disease transmission: application to HIV epidemics among homosexual men in Denmark, Norway and Sweden. *Epidemiology and Infection* **132**(6): 1139-1149.

Anderson, R. M. and R. M. May (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford, Oxford University Press.

Anisimova, M., R. Nielsen and Z. H. Yang (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164**(3): 1229-1236.

Arenas, M. and D. Posada (2007). Recodon: Coalescent simulation of coding DNA sequences with recombination, migration and demography. *Bmc Bioinformatics* **8**(1): 458.

Awadalla, P. (2003). The evolutionary genomics of pathogen recombination. *Nature Reviews Genetics* **4**(1): 50-60.

Ayala, F. J. (1997). Vagaries of the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America* **94**(15): 7776-7783.

Ayres, D. L., A. Darling, D. J. Zwickl, P. Beerli, M. T. Holder, *et al.* (2012). BEAGLE: an Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics. *Systematic Biology* **61**(1):170–173

Baigent, S. J., R. C. Bethell and J. W. McCauley (1999). Genetic analysis reveals that both haemagglutinin and neuraminidase determine the sensitivity of naturally occurring avian influenza viruses to zanamivir in vitro. *Virology* **263**(2): 323-338.

Baigent, S. J. and J. W. McCauley (2001). Glycosylation of haemagglutinin and stalk-length of neuraminidase combine to regulate the growth of avian influenza viruses in tissue culture. *Virus Research* **79**(1-2): 177-185.

Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases*. London and High Wycombe, Charles Griffin and Company Ltd.

Baird, H. A., R. Galetto, Y. Gao, E. Simon-Loriere, M. Abreha, *et al.* (2006a). Sequence determinants of breakpoint location during HIV-1 intersubtype recombination. *Nucleic Acids Research* **34**(18): 5203-5216.

Baird, H.A., Y. Gao, R. Galetto, M. Lalonde, R.M. Anthony *et al*. (2006b). Influence of sequence identity and unique breakpoints on the frequency of intersubtype HIV-1 recombination. *Retrovirology* **3**(91).

Banks, J., E. C. Speidel, J. W. McCauley and D. J. Alexander (2000). Phylogenetic analysis of H7 haemagglutinin subtype influenza A viruses. *Archives of Virology* **145**(5): 1047-1058.

Banks, J., E. S. Speidel, E. Moore, L. Plowright, A. Piccirillo, *et al.* (2001). Changes in the haemagglutinin and the neuraminidase genes prior to the emergence of highly pathogenic H7N1 avian influenza viruses in Italy. *Archives of Virology* **146**(5): 963-973.

Bao, Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zalavsky, *et al.* (2008). The Influenza Virus Resource at the National Center for Biotechnology Information. *J. Virol.* **82**(2): 596-601.

Barré-Sinoussi, F., J. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, *et al.* (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220**(4599): 868-871.

Barton, N. H. and A. M. Etheridge (2004). The effect of selection on genealogies. *Genetics* **166**(2): 1115-1131.

Bataille, A., F. van der Meer, A. Stegeman and G. Koch (2011). Evolutionary Analysis of Inter-Farm Transmission Dynamics in a Highly Pathogenic Avian Influenza Epidemic. *Plos Pathogens* **7**(6).

Bean, W. J., Y. Kawaoka, J. M. Wood, J. E. Pearson and R. G. Webster (1985). Characterization of Virulent and Avirulent a-Chicken-Pennsylvania-83 Influenza-a Viruses - Potential Role of Defective Interfering Rnas in Nature. *Journal of Virology* **54**(1): 151-160.

Becker, W. B. (1966). Isolation and Classification of Tern Virus - Influenza Virus a/Tern/South Africa/1961. *Journal of Hygiene-Cambridge* **64**(3): 309-&.

Beerli, P. and J. Felsenstein (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**(2): 763-773.

Belser, J. A., O. Blixt, L. M. Chen, C. Pappas, T. R. Maines, *et al.* (2008). Contemporary North American influenza H7 viruses possess human receptor specificity: Implications for virus transmissibility. *Proceedings of the National Academy of Sciences of the United States of America* **105**(21): 7558-7563.

Belser, J. A., C. B. Bridges, J. M. Katz and T. M. Tumpey (2009). Past, Present, and Possible Future Human Infection with Influenza Virus A Subtype H7. *Emerging Infectious Diseases* **15**(6): 859-865.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**(1): 289-300.

Blick, T. J., A. Sahasrabudhe, M. McDonald, I. J. Owens, P. J. Morley*, et al.* (1998). The interaction of neuraminidase and hemagglutinin mutations in influenza virus in resistance to 4-guanidino-Neu5Ac2en. *Virology* **246**(1): 95-103.

Bloomquist, E. W. and M. A. Suchard (2010). Unifying Vertical and Nonvertical Evolution: A Stochastic ARG-based Framework. *Systematic Biology* **59**(1): 27-41.

Boender, G. J., T. J. Hagenaars, A. Bouma, G. Nodelijk, A. R. W. Elbers*, et al.* (2007). Risk maps for the spread of highly pathogenic avian influenza in poultry. *PLoS Computational Biology* **3**(4): 704-712.

Bollback, J. P. (2006). SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *Bmc Bioinformatics* **7**.

Boni, M. F., Y. Zhou, J. K. Taubenberger and E. C. Holmes (2008). Homologous recombination is very rare or absent in human influenza A virus. *Journal of Virology* **82**(10): 4807-4811.

Brown, J. D., D. E. Stallknecht, J. R. Beck, D. L. Suarez and D. E. Swayne (2006). Susceptibility of North American ducks and gulls to H5N1 highly pathogenic avian influenza viruses. *Emerging Infectious Diseases* **12**(11): 1663-1670.

Brown, J. D., D. E. Stallknecht and D. E. Swaynet (2008). Experimental infection of swans and geese with highly pathogenic avian influenza virus (H5N1) of Asian lineage. *Emerging Infectious Diseases* **14**(1): 136-142.

Bulach, D., R. Halpin, D. Spiro, L. Pomeroy, D. Janies*, et al.* (2010). Molecular Analysis of H7 Avian Influenza Viruses from Australia and New Zealand: Genetic Diversity and Relationships from 1976 to 2007. *Journal of Virology* **84**(19): 9957-9966.

Burke, D. S. (1997). Recombination in HIV: An important viral evolutionary strategy. *Emerging Infectious Diseases* **3**(3): 253-259.

Bush, R. M., W. M. Fitch, C. A. Bender and N. J. Cox (1999). Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Molecular Biology and Evolution* **16**(11): 1457-1465.

Campitelli, L., E. Mogavero, M. A. De Marco, M. Delogu, S. Puzelli*, et al.* (2004). Interspecies transmission of an H7N3 influenza virus from wild birds to intensively reared domestic poultry in Italy. *Virology* **323**(1): 24-36.

Cardona, C., K. Yee and T. Carpenter (2009). Are live bird markets reservoirs of avian influenza? *Poultry Science* **88**(4): 856-859.

Carvajal-Rodriguez, A., K. A. Crandall and D. Posada (2007). Recombination favors the evolution of drug resistance in HIV-1 during antiretroviral therapy. *Infection Genetics and Evolution* **7**(4): 476-483.

Cavalli-Sforza, L. l. and A. W. F. Edwards (1967). Phylogenetic Analysis - Models and Estimation Procedures. *Evolution* **21**(3): 550-&.

CDC (2004a). CDC Update: Influenza activity - United States and worldwide, 2003–04 season, and compositionof the 2004-05 influenza vaccine. *Morbidity and Mortality Weekly Reports* **53**: 547–552.

CDC (2004b). CDC Update: influenza activity - United States, 2003–04 season. *Morbidity and Mortality Weekly Reports* **53**: 284–287.

CDC (2012). CDC Update: Influenza A (H3N2)v Transmission and Guidelines: Five States, 2011. *Morbidity and Mortality Weekly Reports* **60**(51): 1741-1744.

Chen, H., G. J. D. Smith, S. Y. Zhang, K. Qin, J. Wang, *et al.* (2005). H5N1 virus outbreak in migratory waterfowl. *Nature* **436**(7048): 191-192.

Chen, M.-H. and Q.-M. Shao (1999). Monte Carlo Estimation of Bayesian Credible and HPD Intervals. *Journal of Computational and Graphical Statistics* **8**(1): 69-92.

Chen, R. and E. C. Holmes (2009). Frequent inter-species transmission and geographic subdivision in avian influenza viruses from wild birds. *Virology* **383**(1): 156-161.

Chen, R. B. and E. C. Holmes (2006). Avian influenza virus exhibits rapid evolutionary dynamics. *Molecular Biology and Evolution* **23**(12): 2336-2341.

Chen, W. S., P. A. Calvo, D. Malide, J. Gibbs, U. Schubert, *et al.* (2001). A novel influenza A virus mitochondrial protein that induces cell death. *Nature Medicine* **7**(12): 1306-1312.

Chin, M. P. S., T. D. Rhodes, J. Chen, W. Fu and W. S. Hu (2005). Identification of a major restriction in HIV-1 intersubtype recombination. *Proceedings of the National Academy of Sciences of the United States of America* **102**(25): 9002-9007.

Choi, Y. K., J. H. Lee, G. Erickson, S. M. Goyal, H. S. Joo, *et al.* (2004). H3N2 influenza virus transmission from swine to turkeys, United States. *Emerging Infectious Diseases* **10**(12): 2156-2160.

Choi, Y. K., S. H. Seo, J. A. Kim, R. J. Webby and R. G. Webster (2005). Avian influenza viruses in Korean live poultry markets and their pathogenic potential. *Virology* **332**(2): 529-537.

Claas, E. C. J., A. D. M. E. Osterhaus, R. van Beek, J. C. De Jong, G. F. Rimmelzwaan, *et al.* (1998). Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *Lancet* **351**(9101): 472-477.

Colman, P. M., J. N. Varghese and W. G. Laver (1983). Structure of the Catalytic and Antigenic Sites in Influenza-Virus Neuraminidase. *Nature* **303**(5912): 41-44.

Conenello, G. M., D. Zamarin, L. A. Perrone, T. Tumpey and P. Palese (2007). A single mutation in the PB1-F2 of H5N1 (HK/97) and 1918 influenza A viruses contributes to increased virulence. *Plos Pathogens* **3**(10): 1414-1421.

Connor, R. J., Y. Kawaoka, R. G. Webster and J. C. Paulson (1994). Receptor Specificity in Human, Avian, and Equine H2 and H3 Influenza-Virus Isolates. *Virology* **205**(1): 17-23.

Cottam, E. M., G. Thebaud, J. Wadsworth, J. Gloster, L. Mansley, *et al.* (2008). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society B-Biological Sciences* **275**(1637): 887-895.

Crawford, P. C., E. J. Dubovi, W. L. Castleman, I. Stephenson, E. P. J. Gibbs, *et al.* (2005). Transmission of equine influenza virus to dogs. *Science* **310**(5747): 482-485.

Cunningham, C. W., K. E. Omland and T. H. Oakley (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology & Evolution* **13**(9): 361-366.

Dang, C. C., Q. S. Le, O. Gascuel and S. L. Vinh (2010). FLU, an amino acid substitution model for influenza proteins. *BMC Evolutionary Biology* **10**.

Daniels, R. S., S. Jeffries, P. Yates, G. C. Schild, G. N. Rogers, *et al.* (1987). The Receptor-Binding and Membrane-Fusion Properties of Influenza-Virus Variants Selected Using Anti-Hemagglutinin Monoclonal-Antibodies. *EMBO Journal* **6**(5): 1459-1465.

Dasco, C. C., R. B. Couch, H. R. Six, J. F. Young, J. M. Quaries*, et al.* (1984). Sporadic occurence of zoonotic swine influenza virus infections. *J. Clin. Microb.* **20**: 833-835.

Davison, S., R. J. Eckroade and A. E. Ziegler (2003). A review of the 1996-98 nonpathogenic H7N2 avian influenza outbreak in Pennsylvania. *Avian Diseases* **47**: 823-827.

Davison, S., D. Galligan, T. E. Eckert, A. F. Ziegler and R. J. Eckroade (1999). Economic analysis of an outbreak of avian influenza, 1997-1998. *Journal of the American Veterinary Medical Association* **214**(8): 1164-1167.

Dawood, F. S., S. Jain, L. Finelli, M. W. Shaw, S. Lindstrom*, et al.* (2009). Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans *New England Journal of Medicine* **360**(25): 2605-2615.

Delport, W., A. F. Y. Poon, S. D. W. Frost and S. L. K. Pond (2010). Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**(19): 2455-2457.

Deyde, V. M., X. Y. Xu, R. A. Bright, M. Shaw, C. B. Smith*, et al.* (2007). Surveillance of resistance to adamantanes among influenza A(H3N2) and A(H1N1) viruses isolated worldwide. *Journal of Infectious Diseases* **196**(2): 249-257.

Donnelly, P. (1986). Partition Structures, Polya Urns, the Ewens Sampling Formula and the Ages of Alleles. *Theoretical Population Biology* **30**: 271-288.

Donnelly, P. and S. Tavaré (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29**: 401-421.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips and A. Rambaut (2006). Relaxed phylogenetics and dating with confidence. *Plos Biology* **4**(5): 699-710.

Drummond, A. J., G. K. Nicholls, A. G. Rodrigo and W. Solomon (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**(3): 1307-1320.

Drummond, A. J., O. G. Pybus, A. Rambaut, R. Forsberg and A. G. Rodrigo (2003). Measurably evolving populations. *Trends in Ecology & Evolution* **18**(9): 481-488.

Drummond, A. J. and A. Rambaut (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**.

Drummond, A. J., A. Rambaut, B. Shapiro and O. G. Pybus (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution* **22**(5): 1185-1192.

Duffy, S., L. A. Shackelton and E. C. Holmes (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* **9**(4): 267-276.

Dugan, V. G., R. Chen, D. J. Spiro, N. Sengamalay, J. Zaborsky*, et al.* (2008). The evolutionary genetics and emergence of avian influenza viruses in wild birds. *Plos Pathogens* **4**(5): -.

Dunn, P. A., E. A. Wallner-Pendleton, H. Lu, D. P. Shaw, D. Kradel*, et al.* (2003). Summary of the 2001-02 Pennsylvania H7N2 low pathogenicity avian influenza outbreak in meat type chickens. *Avian Diseases* **47**: 812-816.

Eckert, D. M. and M. S. Kay (2010). Stalking influenza. *Proceedings of the National Academy of Sciences of the United States of America* **107**(31): 13563-13564.

Efron, B. and G. Gong (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *American Statistician* **37**(1): 36-48.

Ekiert, D. C., G. Bhabha, M. A. Elsliger, R. H. E. Friesen, M. Jongeneelen, *et al.* (2009). Antibody Recognition of a Highly Conserved Influenza Virus Epitope. *Science* **324**(5924): 246-251.

Ellis, J., M. Galiano, R. Pebody, A. Lackenby, C. Thompson, *et al.* (2011). Virological analysis of fatal influenza cases in the United Kingdom during the early wave of influenza in winter 2010/11. *Eurosurveillance* **16**(1): 2-7.

Feare, C. J. (2010). Role of Wild Birds in the Spread of Highly Pathogenic Avian Influenza Virus H5N1 and Implications for Global Surveillance. *Avian Diseases* **54**(1): 201-212.

Felsenstein, J. (1973). Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Systematic Zoology* **22**(3): 240-249.

Felsenstein, J. (1981). Evolutionary Trees from DNA-Sequences - a Maximum-Likelihood Approach. *Journal of Molecular Evolution* **17**(6): 368-376.

Felsenstein, J. (1985a). Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. *Evolution* **39**(4): 783-791.

Felsenstein, J. (1985b). Phylogenies and the Comparative Method. *American Naturalist* **125**(1): 1-15.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, Massachusetts, Sinauer Associates.

Ferguson, N. M., D. A. T. Cummings, S. Cauchemez, C. Fraser, S. Riley, *et al.* (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* **437**(7056): 209-214.

Fine, P. E. M. (2003). The Interval between Successive Cases of an Infectious Disease. *American Journal of Epidemiology* **158**(11): 1039-1047.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford, The Clarendon Press.

Fitch, W. M. (1971). Toward Defining Course of Evolution - Minimum Change for a Specific Tree Topology. *Systematic Zoology* **20**(4): 406-&.

Fitch, W. M., J. M. E. Leiter, X. Q. Li and P. Palese (1991). Positive Darwinian Evolution in Human Influenza A Viruses. *Proceedings of the National Academy of Sciences of the United States of America* **88**(10): 4270-4274.

Fouchier, R. A. M., V. Munster, A. Wallensten, T. M. Bestebroer, S. Herfst, *et al.* (2005). Characterization of a novel influenza a virus hemagglutinin subtype (H16) obtained from black-headed gulls. *Journal of Virology* **79**(5): 2814-2822.

Fouchier, R. A. M., P. M. Schneeberger, F. W. Rozendaal, J. M. Broekman, S. A. G. Kemink, *et al.* (2004). Avian influenza A virus (H7N7) associated with human conjunctivitis and a fatal case of acute respiratory distress syndrome. *Proceedings of the National Academy of Sciences of the United States of America* **101**(5): 1356-1361.

Fraser, C., C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, *et al.* (2009). Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings. *Science* **324**(5934): 1557-1561.

Freed, E. O. and M. A. Martin (2007). HIVs and their replication. *Fields' Virology*. D. M. Knipe and P. M. Howley. Philadelphia, Lippincott, Williams, and Wilkins**:** 2107-2186.

Frost, S. D. W. and E. M. Volz (2010). Viral phylodynamics and the search for an 'effective number of infections'. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**(1548): 1879-1890.

Gao, F., E. Bailes, D. L. Robertson, Y. L. Chen, C. M. Rodenburg, *et al.* (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* **397**(6718): 436-441.

Gao, F., L. Yue, A. T. White, P. G. Pappas, J. Barchue, *et al.* (1992). Human Infection by Genetically Diverse Sivsm-Related Hiv-2 in West Africa. *Nature* **358**(6386): 495-499.

Gao, Y., M. Abreha, K. N. Nelson, H. Baird, D. M. Dudley, *et al.* (2011). Enrichment of intersubtype HIV-1 recombinants in a dual infection system using HIV-1 strain-specific siRNAs. *Retrovirology* **8**.

Geraci, J. R., D. J. Staubin, I. K. Barker, R. G. Webster, V. S. Hinshaw, *et al.* (1982). Mass Mortality of Harbor Seals - Pneumonia Associated with Influenza-a Virus. *Science* **215**(4536): 1129-1131.

Gerhard, W., J. Yewdell, M. E. Frankel and R. Webster (1981). Antigenic Structure of Influenza-Virus Hemagglutinin Defined by Hybridoma Antibodies. *Nature* **290**(5808): 713-717.

Geyer, C. J. (1991). *Markov chain Monte Carlo maximum likelihood*. Computing Science and Statistics: Proceedings of the 23rd Symposium of the Interface, Interface Foundation, Fairfax Station, VA.

Gilbert, M., S. H. Newman, J. Y. Takekawa, L. Loth, C. Biradar, *et al.* (2010). Flying Over an Infected Landscape: Distribution of Highly Pathogenic Avian Influenza H5N1 Risk in South Asia and Satellite Tracking of Wild Waterfowl. *Ecohealth* **7**(4): 448-458.

Gohrbandt, S., J. Veits, J. Hundt, J. Bogs, A. Breithaupt, *et al.* (2011). Amino acids adjacent to the haemagglutinin cleavage site are relevant for virulence of avian influenza viruses of subtype H5. *Journal of General Virology* **92**: 51-59.

Goldberg, T. L. (2003). Application of phylogeny reconstruction and character-evolution analysis to inferring patterns of directional microbial transmission. *Preventive Veterinary Medicine* **61**(1): 59-70.

Goldman, N. (1993a). Simple Diagnostic Statistical Tests of Models for DNA Substitution. *Journal of Molecular Evolution* **37**(6): 650-661.

Goldman, N. (1993b). Statistical Tests of Models of DNA Substitution. *Journal of Molecular Evolution* **36**(2): 182-198.

Goldman, N. and Z. H. Yang (1994). Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA-Sequences. *Molecular Biology and Evolution* **11**(5): 725-736.

Gong, J. Z., W. F. Xu and J. Zhang (2007). Structure and functions of influenza virus neuraminidase. *Current Medicinal Chemistry* **14**(1): 113-122.

Gorman, O. T., W. J. Bean, Y. Kawaoka and R. G. Webster (1990a). Evolution of the Nucleoprotein Gene of Influenza-a Virus. *Journal of Virology* **64**(4): 1487-1497.

Gorman, O. T., R. O. Donis, Y. Kawaoka and R. G. Webster (1990b). Evolution of Influenza-a Virus Pb2 Genes - Implications for Evolution of the Ribonucleoprotein Complex and Origin of Human Influenza a Virus. *Journal of Virology* **64**(10): 4893-4902.

Govorkova, E. A., J. E. Rehg, S. Krauss, H. L. Yen, Y. Guan, *et al.* (2005). Lethality to ferrets of H5N1 influenza viruses isolated from humans and poultry in 2004. *Journal of Virology* **79**(4): 2191-2198.

Grassly, N. C. and C. Fraser (2008). Mathematical models of infectious disease transmission. *Nat Rev Micro* **6**(6): 477-487.

Gray, R. R., A. J. Tatem, J. A. Johnson, A. V. Alekseyenko, O. G. Pybus, *et al.* (2011). Testing Spatiotemporal Hypothesis of Bacterial Evolution Using Methicillin-Resistant Staphylococcus aureus ST239 Genome-wide Data within a Bayesian Framework. *Molecular Biology and Evolution* **28**(5): 1593-1603.

Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, *et al.* (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**(5656): 327-332.

Griffiths, R. C. (1981). Transient distribution of the number of segregating sites in a neutral infinite-sites model with no recombination. *J. Appl. Prob.* **18:** 42-51.

Griffiths, R. C. and P. Marjoram (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* **3**(4): 479-502.

Griffiths, R. C. and S. Tavaré (1994). Sampling Theory for Neutral Alleles in a Varying Environment. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **344**(1310): 403-410.

Gubareva, L. V., R. Bethell, G. J. Hart, K. G. Murti, C. R. Penn, *et al.* (1996). Characterization of mutants of influenza A virus selected with the neuraminidase inhibitor 4-guanidino-Neu5Ac2en. *Journal of Virology* **70**(3): 1818-1827.

Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, *et al.* (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**(3): 307-321.

Guo, Y. J., M. Wang, Y. Kawaoka, O. Gorman, T. Ito, *et al.* (1992). Characterization of a New Avian-Like Influenza-a Virus from Horses in China. *Virology* **188**(1): 245-255.

Ha, Y., D. J. Stevens, J. J. Skehel and D. C. Wiley (2002). H5 avian and H9 swine influenza virus haemagglutinin structures: possible origin of influenza subtypes. *Embo Journal* **21**(5): 865-875.

Hahn, B. H., G. M. Shaw, K. M. D. Cock and P. M. Sharp (2000). AIDS as a zoonosis: scientific and public health implications. *Science* **287**: 607-614.

Hale, B. G., R. E. Randall, J. Ortin and D. Jackson (2008). The multifunctional NS1 protein of influenza A viruses. *Journal of General Virology* **89**: 2359-2376.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**: 95-98.

Hancock, K., V. Veguilla, X. H. Lu, W. M. Zhong, E. N. Butler, *et al.* (2009). Cross-Reactive Antibody Responses to the 2009 Pandemic H1N1 Influenza Virus. *New England Journal of Medicine* **361**(20): 1945-1952.

Hasegawa, M., H. Kishino and T. A. Yano (1985). Dating of the Human Ape Splitting by a Molecular Clock of Mitochondrial-DNA. *Journal of Molecular Evolution* **22**(2): 160-174.

Hastie, T., R. Tibshirani and J. Friedman (2001). *Elements of Statistical Learning: Data Mining, Inference and Prediction* New York, Springer-Verlag.

Hastings, W. K. (1970). Monte-Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57**(1): 97-&.

Hatta, M., Y. Hatta, J. H. Kim, S. Watanabe, K. Shinya, *et al.* (2007). Growth of H5N1 influenza a viruses in the upper respiratory tracts of mice. *Plos Pathogens* **3**(10): 1374-1379.

Haydon, D. T., M. E. J. Woolhouse and R. P. Kitching (1997). An analysis of foot-and-mouth-disease epidemics in the UK. *Ima Journal of Mathematics Applied in Medicine and Biology* **14**(1): 1-9.

He, C. Q., G. Z. Han, D. Wang, W. Liu, G. R. Li, *et al.* (2008a). Homologous recombination evidence in human and swine influenza A viruses. *Virology* **380**(1): 12-20.

He, C. Q., Z. H. Xie, G. Z. Han, J. B. Dong and D. Wang (2008b). Homologous recombination as an evolutionary force in the avian influenza A virus. *Molecular Biology and Evolution* **Advanced access**.

Hershberg, R. and D. A. Petrov (2008). Selection on Codon Bias. *Annual Review of Genetics* **42**: 287-299.

Hinshaw, V. S., R. G. Webster, W. J. Bean, J. Downie and D. A. Senne (1983). Swine Influenza Like Viruses in Turkeys - Potential Source of Virus for Humans. *Science* **220**(4593): 206-208.

Hinshaw, V. S., R. G. Webster and B. Turner (1980). The Perpetuation of Orthomyxoviruses and Paramyxoviruses in Canadian Waterfowl. *Canadian Journal of Microbiology* **26**(5): 622-629.

Ho, S. Y. W. and L. S. Jermiin (2004). Tracing the decay of the historical signal in biological sequence data. *Systematic Biology* **53**(4): 623-637.

Holland, J., K. Spindler, F. Horodyski, E. Grabau, S. Nichol, *et al.* (1982). Rapid Evolution of RNA Genomes. *Science* **215**(4540): 1577-1585.

Holmes, E. C. (2004). The phylogeography of human viruses. *Molecular Ecology* **13**(4): 745-756.

Holmes, E. C., E. Ghedin, N. Miller, J. Taylor, Y. M. Bao, *et al.* (2005). Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *Plos Biology* **3**(9): 1579-1589.

Holmes, E. C., S. Nee, A. Rambaut, G. P. Garnett and P. H. Harvey (1995). Revealing the History of Infectious-Disease Epidemics through Phylogenetic Trees. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **349**(1327): 33-40.

Holmes, E. C., L. Q. Zhang, P. Simmonds, A. S. Rogers and A. J. L. Brown (1993). Molecular Investigation of Human-Immunodeficiency-Virus (HIV) Infection in a Patient of an HIV-Infected Surgeon. *Journal of Infectious Diseases* **167**(6): 1411-1414.

Horimoto, T. and Y. Kawaoka (1994). Reverse Genetics Provides Direct Evidence for a Correlation of Hemagglutinin Cleavability and Virulence of an Avian Influenza-a Virus. *Journal of Virology* **68**(5): 3120-3128.

Horimoto, T. and Y. Kawaoka (1995). Molecular-Changes in Virulent Mutants Arising from Avirulent Avian Influenza-Viruses during Replication in 14-Day-Old Embryonated Eggs. *Virology* **206**(1): 755-759.

Horimoto, T. and Y. Kawaoka (2001). Pandemic threat posed by avian influenza A viruses. *Clinical Microbiology Reviews* **14**(1): 129-+.

Hu, W. S. and H. M. Temin (1990). Genetic Consequences of Packaging 2 Rna Genomes in One Retroviral Particle - Pseudodiploidy and High-Rate of Genetic-Recombination. *Proceedings of the National Academy of Sciences of the United States of America* **87**(4): 1556-1560.

Hudson, R. R. (1983). Properties of a Neutral Allele Model with Intragenic Recombination. *Theoretical Population Biology* **23**(2): 183-201.

Hudson, R. R. and N. L. Kaplan (1988). The Coalescent Process in Models with Selection and Recombination. *Genetics* **120**(3): 831-840.

Huelsenbeck, J. P. and K. A. Dyer (2004). Bayesian estimation of positively selected sites. *Journal of Molecular Evolution* **58**(6): 661-672.

Huelsenbeck, J. P., D. M. Hillis and R. Nielsen (1996). A Likelihood-Ratio Test of Monophyly. *Systematic Biology* **45**(4): 546-558.

Huelsenbeck, J. P., R. Nielsen and J. P. Bollback (2003). Stochastic mapping of morphological characters. *Systematic Biology* **52**(2): 131-158.

Huelsenbeck, J. P. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**(8): 754-755.

Hulse-Post, D. J., K. M. Sturm-Ramirez, J. Humberd, P. Seiler, E. A. Govorkova, *et al.* (2005). Role of domestic ducks in the propagation and biological evolution of highly pathogenic H5N1 influenza viruses in Asia. *Proceedings of the National Academy of Sciences of the United States of America* **102**(30): 10682-10687.

Ina, Y. and T. Gojobori (1994). Statistical Analysis of Nucleotide Sequences of the Hemagglutinin Gene of Human Influenza A Viruses. *Proceedings of the National Academy of Sciences of the United States of America* **91**(18): 8388-8392.

ISID. (2004). Avian Influenza - Eastern Asia (28). Retrieved 2011, 7th February, from http://www.promedmail.org/direct.php?id=20040218.0524.

Ito, T., J. N. S. S. Couceiro, S. Kelm, L. G. Baum, S. Krauss, *et al.* (1998). Molecular basis for the generation in pigs of influenza A viruses with pandemic potential. *Journal of Virology* **72**(9): 7367-7373.

Ito, T., K. Okazaki, Y. Kawaoka, A. Takada, R. G. Webster, *et al.* (1995). Perpetuation of Influenza-a Viruses in Alaskan Waterfowl Reservoirs. *Archives of Virology* **140**(7): 1163-1172.

Jackson, D., M. J. Hossain, D. Hickman, D. R. Perez and R. A. Lamb (2008). A new influenza virus virulence determinant: The NS1 protein four C-terminal residues modulate pathogenicity. *Proceedings of the National Academy of Sciences of the United States of America* **105**(11): 4381-4386.

Jeffreys, H. (1961). *Theory of Probability*. 3rd edition. Oxford, Oxford University Press.

Johnson, N. P. A. S. and J. Mueller (2002). Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic. *Bulletin of the History of Medicine* **76**(1): 105-115.

Jonges, M., A. Bataille, R. Enserink, A. Meijer, R. A. M. Fouchier, *et al.* (2011). Comparative Analysis of Avian Influenza Virus Diversity in Poultry and Humans during a Highly Pathogenic Avian Influenza A (H7N7) Virus Outbreak. *Journal of Virology* **85**(20): 10598-10604.

Jukes, T. H. and C. R. Cantor (1969). Evolution of protein molecules. *Mammalian Protein Metabolism*. H. H. Munro. New York, Academic Press. **3**.

Kalish, M. L., K. E. Robbins, D. Pieniazek, A. Schaefer, N. Nzilambi, *et al.* (2004). Recombinant viruses and early global HIV-1 epidemic. *Emerging Infectious Diseases* **10**(7): 1227-1234.

Kaplan, N. L., Darden, T., and Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics* **120**: 819–829.

Kashyap, A. K., J. Steel, A. F. Oner, M. A. Dillon, R. E. Swale, *et al.* (2008). Combinatorial antibody libraries from survivors of the Turkish H5N1 avian influenza outbreak reveal virus neutralization strategies. *Proceedings of the National Academy of Sciences of the United States of America* **105**(16): 5986-5991.

Kass, R. E. and A. E. Raftery (1995). Bayes Factors. *Journal of the American Statistical Association* **90**(430): 773-795.

Katz, J. M., X. H. Lu, T. M. Tumpey, C. B. Smith, M. W. Shaw, *et al.* (2000). Molecular correlates of influenza A H5N1 virus pathogenesis in mice. *Journal of Virology* **74**(22): 10807-10810.

Kaverin, N. V., A. S. Gambaryan, N. V. Bovin, I. A. Rudneva, A. A. Shilov, *et al.* (1998). Postreassortment changes in influenza A virus hemagglutinin restoring HA-NA functional match. *Virology* **244**(2): 315-321.

Kaverin, N. V., M. N. Matrosovich, A. S. Gambaryan, I. A. Rudneva, A. A. Shilov, *et al.* (2000). Intergenic HA-NA interactions in influenza A virus: postreassortment substitutions of charged amino acid in the hemagglutinin of different subtypes. *Virus Research* **66**(2): 123-129.

Kawaoka, Y., S. Krauss and R. G. Webster (1989). Avian-to-Human Transmission of the Pb1 Gene of Influenza-a Viruses in the 1957 and 1968 Pandemics. *Journal of Virology* **63**(11): 4603-4608.

Kenah, E., M. Lipsitch and J. M. Robins (2008). Generation interval contraction and epidemic data analysis. *Mathematical Biosciences* **213**(1): 71-79.

Kermack, W. O. and A. G. McKendrick (1927). Contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character* **115**(772): 700-721.

Kida, H., T. Ito, J. Yasuda, Y. Shimizu, C. Itakura, *et al.* (1994). Potential for Transmission of Avian Influenza-Viruses to Pigs. *Journal of General Virology* **75**: 2183-2188.

Kimura, M. (1968). Genetic Variability Maintained in a Finite Population Due to Mutational Production of Neutral and Nearly Neutral Isoalleles. *Genetical Research* **11**(3): 247-&.

Kimura, M. (1980). A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide-Sequences. *Journal of Molecular Evolution* **16**(2): 111-120.

Kimura, M. (1986). DNA and the Neutral Theory. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **312**(1154): 343-354.

Kingman, J. (1982). The Coalescent. *Stochastic Processes and their Applications* **13**: 235-248.

Kongchanagul, A., O. Suptawiwat, P. Kanrai, M. Uiprasertkul, P. Puthavathana, *et al.* (2008). Positive selection at the receptor-binding site of haemagglutinin H5 in viral sequences derived from human tissues. *Journal of General Virology* **89**: 1805-1810.

Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, *et al.* (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**(5472): 1789-1796.

Kosakovsky Pond, S., A. F. Poon and S. D. Frost (2009). Estimating selection pressures on alignments of coding sequences. *The Phylogenetic Handbook*. Second edition. P. Lemey, M. Salemi and A. M. Vandamme. Cambridge, Cambridge University Press**:** 419-490.

Kosakovsky Pond, S. L. and S. D. W. Frost (2005). Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution* **22**(5): 1208-1222.

Kosakovsky Pond, S. L., S. D. W. Frost and S. V. Muse (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**(5): 676-679.

Kosakovsky Pond, S. L., A. F. Y. Poon, A. J. L. Brown and S. D. W. Frost (2008). A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza a virus. *Molecular Biology and Evolution* **25**(9): 1809-1824.

Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk and S. D. W. Frost (2006a). Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular Biology and Evolution* **23**(10): 1891-1901.

Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk and S. D. W. Frost (2006b). GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22**(24): 3096-3098.

Krasnitz, M., A. J. Levine and R. Rabadan (2008). Anomalies in the influenza virus genome database: New biology or laboratory errors? *Journal of Virology* **82**(17): 8947-8950.

Krauss, S., D. Walker, S. P. Pryor, L. Niles, C. H. Li, *et al.* (2004). Influenza A viruses of migrating wild aquatic birds in North America. *Vector-Borne and Zoonotic Diseases* **4**(3): 177-189.

Krone, S. M., and Neuhauser, C. (1997). Ancestral processes with selection. *Theor. Pop. Biol.* **51**: 210–237.

Krystal, M., R. M. Elliott, E. W. Benz, J. F. Young and P. Palese (1982). Evolution of Influenza-a and Influenza-B Viruses - Conservation of Structural Features in the Hemagglutinin Genes. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* **79**(15): 4800-4804.

Kuhner, M. K., J. Yamato and J. Felsenstein (1995). Estimating Effective Population Size and Mutation Rate from Sequence Data Using Metropolis-Hastings Sampling. *Genetics* **140**(4): 1421-1430.

Kuiken, T., E. C. Holmes, J. McCauley, G. F. Rimmelzwaan, C. S. Williams, *et al.* (2006). Host species barriers to influenza virus infections. *Science* **312**(5772): 394-397.

Kurtz, J., R. J. Manvell and J. Banks (1996). Avian influenza virus isolated from a woman with conjunctivitis. *Lancet* **348**(9031): 901-902.

Lebarbenchon, C. and D. E. Stallknecht (2011). Host shifts and molecular evolution of H7 avian influenza virus hemagglutinin. *Virology Journal* **8**.

Lemey, P., O. G. Pybus, A. Rambaut, A. J. Drummond, D. L. Robertson, *et al.* (2004). The molecular population genetics of HIV-1 group O. *Genetics* **167**(3): 1059-1068.

Lemey, P., A. Rambaut, A. J. Drummond and M. A. Suchard (2009). Bayesian Phylogeography Finds Its Roots. *PLoS Computational Biology* **5**(9).

Lewis, F., G. J. Hughes, A. Rambaut, A. Pozniak and A. J. L. Brown (2008). Episodic sexual transmission of HIV revealed by molecular phylodynamics. *Plos Medicine* **5**(3): 392-402.

Li, K. S., Y. Guan, J. Wang, G. J. D. Smith, K. M. Xu, *et al.* (2004). Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature* **430**(6996): 209-213.

Li, M. and B. Wang (2007). Homology modeling and examination of the effect of the D92E mutation on the H5N1 nonstructural protein NS1 effector domain. *Journal of Molecular Modeling* **13**: 1237–1244.

Li, W. H., C. I. Wu and C. C. Luo (1985). A New Method for Estimating Synonymous and Nonsynonymous Rates of Nucleotide Substitution Considering the Relative Likelihood of Nucleotide and Codon Changes. *Molecular Biology and Evolution* **2**(2): 150-174.

Lipsitch, M. and C. T. Bergstrom (2004). Invited commentary: Real-time tracking of control measures for emerging infections. *American Journal of Epidemiology* **160**(6): 517-519.

Liu, J., H. Xiao, F. Lei, Q. Zhu, K. Qin, *et al.* (2005). Highly pathogenic H5N1 influenza virus infection in migratory birds. *Science* **309**(5738): 1206-1206.

Liu, M., S. Q. He, D. Walker, N. N. Zhou, D. R. Perez, *et al.* (2003). The influenza virus gene pool in a poultry market in South Central China. *Virology* **305**(2): 267-275.

Liu, S. L., J. E. Mittler, D. C. Nickle, T. M. Mulvania, D. Shriner, *et al.* (2002). Selection for human immunodeficiency virus type 1 recombinants in a patient with rapid progression to AIDS. *Journal of Virology* **76**(21): 10674-10684.

Ludwig, S., L. Stitz, O. Planz, H. Van, W. M. Fitch, *et al.* (1995). European Swine Virus as a Possible Source for the Next Influenza Pandemic. *Virology* **212**(2): 555-561.

Lycett, S. J., G. Baillie, E. Coulter, S. Bhatt, P. Kellam, *et al.* (2012). Estimating reassortment rates in co-circulating Eurasian swine influenza viruses. *Journal of General Virology (in press).*

Lycett, S. J., M. J. Ward, F. I. Lewis, A. F. Y. Poon, S. L. K. Pond, *et al.* (2009). Detection of Mammalian Virulence Determinants in Highly Pathogenic Avian Influenza H5N1 Viruses: Multivariate Analysis of Published Data. *Journal of Virology* **83**(19): 9901-9910.

Macken, C. A., R. J. Webby and W. J. Bruno (2006). Genotype turnover by reassortment of replication complex genes from avian Influenza A virus. *Journal of General Virology* **87**: 2803-2815.

Maddison, W. P. (1990). A Method for Testing the Correlated Evolution of 2 Binary Characters - Are Gains or Losses Concentrated on Certain Branches of a Phylogenetic Tree. *Evolution* **44**(3): 539-557.

Maddison, W. P. (1995). Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. *Systematic Biology* **44**(4): 474-481.

Maddison, W. P. and D. R. Maddison (1992). *MacClade: Analysis of phylogeny and character evolution. Version 3.* Sunderland, Massachusetts, Sinauer Associates.

Makarova, N. A., N. V. Kaverin, S. Krauss, D. Senne and R. G. Webster (1999). Transmission of Eurasian avian H2 influenza virus to shorebirds in North America. *Journal of General Virology* **80**: 3167-3171.

Mannelli, A., L. Busani, M. Toson, S. Bertolini and S. Marangon (2007). Transmission parameters of highly pathogenic avian influenza (H7N1) among industrial poultry farms in northern Italy in 1999-2000. *Preventive Veterinary Medicine* **81**(4): 318-322.

Mannelli, A., N. Ferre and S. Marangon (2006). Analysis of the 1999-2000 highly pathogenic avian influenza (H7N1) epidemic in the main poultry-production area in northern Italy. *Preventive Veterinary Medicine* **73**(4): 273-285.

Margush, T. and F. R. McMorris (1981). Consensus N-Trees. *Bulletin of Mathematical Biology* **43**(2): 239-244.

Matrosovich, M., A. Tuzikov, N. Bovin, A. Gambaryan, A. Klimov, *et al.* (2000). Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. *Journal of Virology* **74**(18): 8502-8512.

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

Matrosovich, M. N., A. S. Gambaryan, S. Teneberg, V. E. Piskarev, S. S. Yamnikova, *et al.* (1997). Avian influenza a viruses differ from human viruses by recognition of sialyloligosaccharides and gangliosides and by a higher conservation of the HA receptor-binding site. *Virology* **233**(1): 224-234.

Matrosovich, M. N., S. Krauss and R. G. Webster (2001). H9N2 influenza a viruses from poultry in Asia have human virus-like receptor specificity. *Virology* **281**(2): 156-162.

Matthews, L. and M. Woolhouse (2005). New approaches to quantifying the spread of infection. *Nature Reviews Microbiology* **3**(7): 529-536.

McKimm-Breschkin, J. L., A. Sahasrabudhe, T. J. Blick, M. McDonald, P. M. Colman, *et al.* (1998). Mutations in a conserved residue in the influenza virus neuraminidase active site decreases sensitivity to Neu5Ac2en-derived inhibitors. *Journal of Virology* **72**(3): 2456-2462.

McKimmBreschkin, J. L., T. J. Blick, A. Sahasrabudhe, T. Tiong, D. Marshall, *et al.* (1996). Generation and characterization of variants of NWS/G70C influenza virus after in vitro passage in 4-amino-Neu5Ac2en and 4-guanidino-Neu5Ac2en. *Antimicrobial Agents and Chemotherapy* **40**(1): 40-46.

McQuiston, J. H., L. P. Garber, B. A. Porter-Spalding, J. W. Hahn, F. W. Pierson, *et al.* (2005). Evaluation of risk factors for the spread of low pathogenicity H7N2 avian influenza virus among commercial poultry farms. *Javma-Journal of the American Veterinary Medical Association* **226**(5): 767-772.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **21**(6): 1087-1092.

Minin, V. N., E. W. Bloomquist and M. A. Suchard (2008). Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. *Molecular Biology and Evolution* **25**(7): 1459-1471.

Minin, V. N., K. S. Dorman, F. Fang and M. A. Suchard (2007). Phylogenetic Mapping of Recombination Hotspots in Human Immunodeficiency Virus via Spatially Smoothed Change-Point Processes. *Genetics* **175**(4): 1773-1785.

Minin, V. N. and M. A. Suchard (2008a). Counting labeled transitions in continuous-time Markov models of evolution. *Journal of Mathematical Biology* **56**(3): 391-412.

Minin, V. N. and M. A. Suchard (2008b). Fast, accurate and simulation-free stochastic mapping. *Philosophical Transactions of the Royal Society B-Biological Sciences* **363**(1512): 3985-3995.

Miyata, T. (1984). Evolution of DNA: dynamically evolving eukaryotic genes. *Introduction to molecular evolutionary study. (In Japanese).* M. Kimura. Tokyo, Baifukan**:** 56-90.

Miyata, T. and T. Yasunaga (1980). Molecular Evolution of Messenger-RNA - a Method for Estimating Evolutionary Rates of Synonymous and Amino-Acid Substitutions from Homologous Nucleotide-Sequences and Its Application. *Journal of Molecular Evolution* **16**(1): 23-36.

Morens, D. M., J. K. Taubenberger and A. S. Fauci (2010). The 2009 H1N1 Pandemic Influenza Virus: What Next? *Mbio* **1**(4).

Moscona, A. (2009). Global Transmission of Oseltamivir-Resistant Influenza. *New England Journal of Medicine* **360**(10): 953-956.

Moss, W. N., S. F. Priore and D. H. Turner (2011). Identification of potential conserved RNA secondary structure throughout influenza A coding regions. *RNA - A Publication of the RNA Society* **17**(6): 991-1011.

Mulanga-Kabeya, C., N. Nzilambi, B. Edidi, M. Minlangu, T. Tshimpaka*, et al.* (1998). Evidence of stable HIV seroprevalences in selected populations in the Democratic Republic of the Congo. *Aids* **12**(8): 905-910.

Munster, V. J., C. Baas, P. Lexmond, J. Waldenstrom, A. Wallensten*, et al.* (2007). Spatial, temporal, and species variation in prevalence of influenza A viruses in wild migratory birds. *Plos Pathogens* **3**(5): 630-638.

Muse, S. V. and B. S. Gaut (1994). A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome. *Molecular Biology and Evolution* **11**(5): 715-724.

Nagarajan, N. and C. Kingsford (2011). GiRaF: robust, computational identification of influenza reassortments via graph mining. *Nucleic Acids Research* **39**(6): e34.

Nakajima, K., S. Nakajima and A. Sugiura (1982). The Possible Origin of H3n2 Influenza-Virus. *Virology* **120**(2): 504-509.

NCBI. (2012). Growth of Influenza Virus Sequences in GenBank. Retrieved 27 April, 2012, from http://www.ncbi.nlm.nih.gov/genomes/FLU/growth.html.

Nee, S., E. C. Holmes, A. Rambaut and P. H. Harvey (1995). Inferring Population History from Molecular Phylogenies. *Philosophical Transactions: Biological Sciences* **349**(1327): 25-31.

Nei, M. and T. Gojobori (1986). Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions. *Molecular Biology and Evolution* **3**(5): 418-426.

Nelson, M. I. and E. C. Holmes (2007). The evolution of epidemic influenza. *Nature Reviews Genetics* **8**(3): 196-205.

Nelson, M. I., C. Viboud, L. Simonsen, R. T. Bennett, S. B. Griesemer*, et al.* (2008). Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. *PLoS Pathogens* **4**(2).

Nestorowicz, A., Y. Kawaoka, W. J. Bean and R. G. Webster (1987). Molecular Analysis of the Hemagglutinin Genes of Australian H7n7 Influenza-Viruses - Role of Passerine Birds in Maintenance or Transmission. *Virology* **160**(2): 411-418.

Neuhauser, C., and Krone, S. M. (1997). The genealogy of samples in models with selection. *Genetics* **145**: 519–534.

Newton, M. A., A. E. Raftery, A. C. Davison, M. Bacha, G. Celeux*, et al.* (1994). Approximate Bayesian-Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society Series B-Methodological* **56**(1): 3-48.

Nickbakhsh, S., L. Matthews, P. R. Bessell, S. W. J. Reid and R. R. Kao (2011). Generating social network data using partially described networks: an example informing avian influenza control in the British poultry industry. *Bmc Veterinary Research* **7**.

Nickle, D. C., L. Heath, M. A. Jensen, P. B. Gilbert, J. I. Mullins*, et al.* (2007). HIV-Specific Probabilistic Models of Protein Evolution. *Plos One* **2**(6).

Nielsen, R. (2001). Mutations as missing data: Inferences on the ages and distributions of nonsynonymous and synonymous mutations. *Genetics* **159**(1): 401-411.

Nielsen, R. (2002). Mapping mutations on phylogenies. *Systematic Biology* **51**(5): 729-739.

Nielsen, R. and J. P. Huelsenbeck (2002). *Detecting positively selected amino acid sites using posterior predictive p-values*. Pacific Symposium on Biocomputing, Proceedings, Singapore, World Scientific.

Nielsen, R. and Z. H. Yang (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**(3): 929-936.

Nobusawa, E., T. Aoyama, H. Kato, Y. Suzuki, Y. Tateno, *et al.* (1991). Comparison of Complete Amino-Acid-Sequences and Receptor-Binding Properties among 13 Serotypes of Hemagglutinins of Influenza a-Viruses. *Virology* **182**(2): 475-485.

Nordborg, M. (1999). The coalescent with partial selfing and balancing selection: An application of structured coalescent processes. *Statistics in Molecular Biology and Genetics, IMS Lecture Notes-Monograph Series.* F. Seillier-Moiseiwitsch, ed. Institute of Mathematical Statistics, Hayward, California, **33**: 56–76.

Nordborg, M. (2001). Coalescent Theory. *Handbook of Statistical Genetics.* D. J. Balding, M.J. Bishop, & C. Cannings. John Wiley & Sons, Chichester, UK: 179–212.

O'Brien, J. D., V. N. Minin and M. A. Suchard (2009). Learning to Count: Robust Estimates for Labeled Distances between Molecular Sequences. *Molecular Biology and Evolution* **26**(4): 801-814.

Obenauer, J. C., J. Denson, P. K. Mehta, X. P. Su, S. Mukatira, *et al.* (2006). Large-scale sequence analysis of avian influenza isolates. *Science* **311**(5767): 1576-1580.

Okuno, Y., Y. Isegawa, F. Sasao and S. Ueda (1993). A Common Neutralizing Epitope Conserved between the Hemagglutinins of Influenza-a Virus H1 and H2 Strains. *Journal of Virology* **67**(5): 2552-2558.

Olsen, B., V. J. Munster, A. Wallensten, J. Waldenstrom, A. D. M. E. Osterhaus, *et al.* (2006). Global patterns of influenza A virus in wild birds. *Science* **312**(5772): 384-388.

Olusa, T. A. O., A. K. Adegunwa, A. A. Aderonmu and C. A. O. Adeyefa (2010). Serological Evidence Of Equine H7 Influenza Virus in Polo Horses in Nigeria. *Science World Journal* **5**(2): 17-19.

Onafuwa-Nuga, A. and A. Telesnitsky (2009). The Remarkable Frequency of Human Immunodeficiency Virus Type 1 Genetic Recombination. *Microbiol. Mol. Biol. Rev.* **73**(3): 451-480.

Opgen-Rhein, R., L. Fahrmeir and K. Strimmer (2005). Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *Bmc Evolutionary Biology* **5**.

Pagel, M. (1994). Detecting Correlated Evolution on Phylogenies - a General-Method for the Comparative-Analysis of Discrete Characters. *Proceedings of the Royal Society of London Series B-Biological Sciences* **255**(1342): 37-45.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature* **401**(6756): 877-884.

Pagel, M. and A. Meade (2006). Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *American Naturalist* **167**(6): 808-825.

Paillart, J. C., M. Dettenhofer, X. F. Yu, C. Ehresmann, B. Ehresmann, *et al.* (2004). First snapshots of the HIV-1 RNA structure in infected cells and in virions. *J. Biol. Chem.* **279**(46): 48397-48403.

Palese, P. (1977). Genes of Influenza-Virus. *Cell* **10**(1): 1-10.

Palese, P. and M. L. Shaw (2007). Orthomyxoviridae: the viruses and their replication. *Fields' Virology***:** 1647-1689.

Palese, P., K. Tobita, M. Ueda and R. W. Compans (1974). Characterization of Temperature Sensitive Influenza-Virus Mutants Defective in Neuraminidase. *Virology* **61**(2): 397-410.

Panigrahy, B. and D. A. Senne (2003). Subtypes of Avian Influenza Virus Isolated from Exotic Birds and Ratites in the United States, 1992-1996. *Avian Diseases* **47**(ArticleType: research-article / Issue Title: Special Issue, Fourth International Symposium on Avian Influenza, 1997 Proceedings / Full publication date: 2003 / Copyright © 2003 American Association of Avian Pathologists): 70-75.

Panigrahy, B., D. A. Senne and J. C. Pedersen (2002). Avian influenza virus subtypes inside and outside the live bird markets, 1993-2000: A spatial and temporal relationship. *Avian Diseases* **46**(2): 298-307.

Pappas, C., Y. Matsuoka, D. E. Swayne and R. O. Donis (2007). Development and evaluation of an influenza virus subtype H7N2 vaccine candidate for pandemic preparedness. *Clinical and Vaccine Immunology* **14**(11): 1425-1432.

Parker, J., A. Rambaut and O.G. Pybus (2008). Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infection, Genetics and Evolution* **8**:239-46

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence* **29**: 241–288.

Pearson, J. E. (2003). International standards for the control of avian influenza. *Avian Diseases* **47**: 972-975.

Peiris, J. S. M., M. D. de Jong and Y. Guan (2007). Avian influenza virus (H5N1): a threat to human health. *Clinical Microbiology Reviews* **20**(2): 243-267.

Peiris, J. S. M., Y. Guan, D. Markwell, P. Ghose, R. G. Webster, *et al.* (2001). Cocirculation of avian H9N2 and contemporary "human" H3N2 influenza A viruses in pigs in southeastern China: Potential for genetic reassortment? *Journal of Virology* **75**(20): 9679-9686.

Perdue, M. L., M. Garcia and D. Senne (1997). Virulence-associated sequence duplication at the hemagglutinin cleavage site of avian influenza viruses. *Virus Research* **49**(2): 173-186.

Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard and D. D. Ho (1996). HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**(5255): 1582-1586.

Perler, F., A. Efstratiadis, P. Lomedico, W. Gilbert, R. Kolodner, *et al.* (1980). The Evolution of Genes - the Chicken Preproinsulin Gene. *Cell* **20**(2): 555-566.

Philippe, H., F. Delsuc, H. Brinkmann and N. Lartillot (2005). Phylogenomics. *Annual Review of Ecology Evolution and Systematics* **36**: 541-562.

Pilcher, C. D., J. K. Wong and S. K. Pillai (2008). Inferring HIV transmission dynamics from phylogenetic sequence relationships. *Plos Medicine* **5**(3): 350-352.

Pinto, L. H. and R. A. Lamb (1995). Understanding the Mechanism of Action of the Anti-influenza Virus Drug Amantadine. *Trends in Microbiology* **3**(7): 271-271.

Piyasirisilp, S., F. E. McCutchan, J. K. Carr, E. Sanders-Buell, W. Liu, *et al.* (2000). A recent outbreak of human immunodeficiency virus type 1 infection in southern China was initiated by two highly homogeneous, geographically separated strains, circulating recombinant form AE and a novel BC recombinant. *Journal of Virology* **74**(23): 11286-11295.

Plantier, J.-C., M. Leoz, J. E. Dickerson, F. De Oliveira, F. Cordonnier, *et al.* (2009). A new human immunodeficiency virus derived from gorillas. *Nature Medicine* **15**(8): 871-872.

Posada, D. and K. A. Crandall (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**(9): 817-818.

Posada, D. and K. A. Crandall (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* **98**(24): 13757-13762

Posada, D. and K. A. Crandall (2002). The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution* **54**: 396-402.

Preston, B. D., B. J. Poiesz and L. A. Loeb (1988). Fidelity of HIV-1 Reverse-Transcriptase. *Science* **242**(4882): 1168-1171.

Pybus, O. G., M. A. Charleston, S. Gupta, A. Rambaut, E. C. Holmes, *et al.* (2001). The epidemic behavior of the hepatitis C virus. *Science* **292**(5525): 2323-2325.

Pybus, O. G. and A. Rambaut (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics* **10**(8): 540-550.

Pybus, O. G., A. Rambaut and P. H. Harvey (2000). An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies. *Genetics* **155**(3): 1429-1437.

Raghwani, J., A. Rambaut, E. C. Holmes, V. T. Hang, T. T. Hien, *et al.* (2011). Endemic Dengue Associated with the Co-Circulation of Multiple Viral Lineages and Localized Density-Dependent Transmission. *Plos Pathogens* **7**(6).

Rambaut, A. and A. J. Drummond (2007). Tracer v1.4. Available from http://beast.bio.ed.ac.uk/Tracer.

Rambaut, A., D. Posada, K. A. Crandall and E. C. Holmes (2004). The causes and consequences of HIV evolution. *Nature Reviews Genetics* **5**(1): 52-61.

Rambaut, A., O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger, *et al.* (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453**(7195): 615-U612.

Rambaut, A., D. L. Robertson, O. G. Pybus, M. Peeters and E. C. Holmes (2001). Human immunodeficiency virus - Phylogeny and the origin of HIV-1. *Nature* **410**(6832): 1047-1048.

Ratner, L., W. Haseltine, R. Patarca, K. J. Livak, B. Starcich, *et al.* (1985). Complete Nucleotide-Sequence of the Aids Virus, Htlv-Iii. *Nature* **313**(6000): 277-284.

Reid, A. H., T. G. Fanning, J. V. Hultin and J. K. Taubenberger (1999). Origin and evolution of the 1918 "Spanish" influenza virus hemagglutinin gene. *Proceedings of the National Academy of Sciences of the United States of America* **96**(4): 1651-1656.

Reid, A. H. and J. K. Taubenberger (2003). The origin of the 1918 pandemic influenza virus: a continuing enigma. *Journal of General Virology* **84**: 2285-2292.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambirdge University Press.

Roberts, J. D., K. Bebenek and T. A. Kunkel (1988). The Accuracy of Reverse Transcriptase from HIV-1. *Science* **242**(4882): 1171-1173.

Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, *et al.* (2000). HIV-1 nomenclature proposal. *Science* **288**(5463): 55-57.

Robertson, D. L., P. M. Sharp, F. E. Mccutchan and B. H. Hahn (1995). Recombination in HIV-1. *Nature* **374**(6518): 124-126.

Rodrigo, A. G. (2009). The coalescent: population genetic inference using genealogies. *The Phylogenetic Handbook*. Second edition. P. Lemey, M. Salemi and A. M. Vandamme. Cambridge, Cambridge Universitt Press**:** 549-561.

Rogers, G. N. and J. C. Paulson (1983). Receptor Determinants of Human and Animal Influenza-Virus Isolates - Differences in Receptor Specificity of the Hemagglutinin-H-3 Based on Species of Origin. *Virology* **127**(2): 361-373.

Rohm, C., T. Horimoto, Y. Kawaoka, J. Suss and R. G. Webster (1995). Do Hemagglutinin Genes of Highly Pathogenic Avian Influenza-Viruses Constitute Unique Phylogenetic Lineages. *Virology* **209**(2): 664-670.

Ronquist, F. and J. P. Huelsenbeck (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**(12): 1572-1574.

Rott, R. (1980). Genetic-Determinants for Infectivity and Pathogenicity of Influenza-Viruses. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **288**(1029): 393-&.

Rott, R., M. Orlich and C. Scholtissek (1976). Attenuation of Pathogenicity of Fowl Plague Virus by Recombination with Other Influenza a Viruses Nonpathogenic for Fowl - Non-Exclusive Dependence of Pathogenicity on Hemagglutinin and Neuraminidase of Virus. *Journal of Virology* **19**(1): 54-60.

Rudneva, I. A., N. A. Il'yushina, A. A. Shilov, N. L. Varich, B. V. Sinitsyn*, et al.* (2003). Functional interactions of the influenza virus glycoproteins. *Molecular Biology* **37**(1): 31-36.

Russell, R. J., S. J. Gamblin, L. F. Haire, D. J. Stevens, B. Xiao*, et al.* (2004). HI and H7 influenza haemagglutinin structures extend a structural classification of haemagglutinin subtypes. *Virology* **325**(2): 287-296.

Russell, R. J., L. F. Haire, D. J. Stevens, P. J. Collins, Y. P. Lin*, et al.* (2006). The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **443**(7107): 45-49.

Saitou, N. and M. Nei (1987). The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution* **4**(4): 406-425.

Salminen, M. and D. Martin (2009). Detecting and characterising individual recombination events. *The Phylogenetic Handbook*. Second. P. Lemey, M. Salemi and A. M. Vandamme. Cambridge, Cambridge University Press**:** 519-548.

Schafer, J. R., Y. Kawaoka, W. J. Bean, J. Suss, D. Senne*, et al.* (1993). Origin of the Pandemic 1957 H2 Influenza-a Virus and the Persistence of Its Possible Progenitors in the Avian Reservoir. *Virology* **194**(2): 781-788.

Schafer, W. (1955). Vergleichende Sero-Immunologische Untersuchungen Uber Die Viren Der Influenza Und Klassischen Geflugelpest. *Zeitschrift Fur Naturforschung Part B-Chemie Biochemie Biophysik Biologie Und Verwandten Gebiete* **10**(2): 81-91.

Schierup, M. and J. Hein (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**(897-891).

Schierup, M. H. and R. Forsberg (2003). *Recombination and phylogenetic analysis of HIV-1*. Origin of HIV and emerging persistent viruses, Accademia Nazionale dei Lincei, Rome.

Schluter, D., T. Price, A. O. Mooers and D. Ludwig (1997). Likelihood of ancestor states in adaptive radiation. *Evolution* **51**(6): 1699-1711.

Scholtissek, C., H. Burger, O. Kistner and K. F. Shortridge (1985). The Nucleoprotein as a Possible Major Factor in Determining Host Specificity of Influenza H3n2 Viruses. *Virology* **147**(2): 287-294.

Scholtissek, C., W. Rohde, V. Vonhoyningen and R. Rott (1978). Origin of Human Influenza-Virus Subtypes H2n2 and H3n2. *Virology* **87**(1): 13-20.

Schwartz, G. (1979). Estimating the dimension of a model. *Annals of Statistics* **6**: 461-464.

Senne, D. A., B. Panigrahy, Y. Kawaoka, J. E. Pearson, J. Süss, *et al.* (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 Diseases* **40**(2): 425-437.

Senne, D. A., J. C. Pedersen and B. Panigrahy. (2003a). Live-bird markets in the Northeastern United States: a source of avian influenza in commercial poultry. Retrieved 15th February, 2012, from http://birdflubook.com/resources/senne19.pdf.

Senne, D. A., D. L. Suarez, J. C. Pedersen and B. Panigrahy (2003b). Molecular and biological characteristics of H5 and H7 avian influenza viruses in live-bird markets of the northeastern United States 1994-2001. *Avian Diseases* **47**: 898-904.

Seo, S. H., E. Hoffmann and R. G. Webster (2002). Lethal H5N1 influenza viruses escape host anti-viral cytokine responses. *Nature Medicine* **8**(9): 950-954.

Seto, J. T. and R. Rott (1966). Functional Significance of Sialidase during Influenza Virus Multiplication. *Virology* **30**(4): 731-&.

Shackelton, L. A., C. R. Parrish, U. Truyen and E. C. Holmes (2005). High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proceedings of the National Academy of Sciences of the United States of America* **102**(2): 379-384.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, *et al.* (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**(11): 2498-2504.

Shapiro, B., A. Rambaut and A. J. Drummond (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology and Evolution* **23**(1): 7-9.

Sharp, P. M., E. Bailes, R. R. Chaudhuri, C. M. Rodenburg, M. O. Santiago, *et al.* (2001). The origins of acquired immune deficiency syndrome viruses: where and when? *Philosophical Transactions of the Royal Society B-Biological Sciences* **356**(1410): 867-876.

Sharp, P. M., L. R. Emery and K. Zeng (2010). Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**(1544): 1203-1212.

Sharp, P. M. and B. H. Hahn (2008). Prehistory of HIV-1. *Nature* **455**: 605-606.

Sharp, P. M. and B. H. Hahn (2011). Origins of HIV and the AIDS pandemic. *Cold Spring Harbor Perspectives in Medicine* **1**.

Shimodaira, H. and M. Hasegawa (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* **16**(8): 1114-1116.

Shriner, D., D. C. Nickle, M. A. Jensen and J. I. Mullins (2003). Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genetical Research* **81**(2): 115-121.

Simmonds, P. (2012). SSE: a nucleotide and amino acid sequence analysis platform. *BMC Research Notes* **5**(50).

Simmonds, P. and S. Midgley (2005). Recombination in the genesis and evolution of hepatitis B virus genotypes. *J. Virol.* **79**(24): 15467-15476.

Simmonds, P. and D. B. Smith (1999). Structural constraints on RNA virus evolution. *Journal of Virology* **73**(7): 5787-5794.

Sjodin, P., I. Kaj, S. Krone, M. Lascoux and M. Nordborg (2005). On the meaning and existence of an effective population size. *Genetics* **169**(2): 1061-1070.

Skehel, J. J. and D. C. Wiley (2000). Receptor binding and membrane fusion in virus entry: The influenza hemagglutinin. *Annual Review of Biochemistry* **69**: 531-569.

Sloth Andersen, E., R. E. Jeeninga, C. K. Damgaard, B. Berkhout and J. Kjems (2003). Dimerization and template switching in the 5 ' untranslated region between various subtypes of human immunodeficiency virus type 1. *J. Virol.* **77**(5): 3020-3030.

Smith, B. (2007). boa: An R Package for MCMC Output Convergence Assessment and Posterior Inference. *Journal of Statistical Software* **21**(11).

Smith, G. J. D., D. Vijaykrishna, J. Bahl, S. J. Lycett, M. Worobey, *et al.* (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**(7250): 1122-U1107.

Smith, W., C. H. Andrewes and P. P. Laidlaw (1933). A virus obtained from influenza patients. *Lancet* **2**: 66-68.

Spackman, E., K. G. McCracken, K. Winker and D. E. Swayne (2006). H7N3 avian influenza virus found in a South American wild duck is related to the Chilean 2002 poultry outbreak, contains genes from equine and north American wild bird lineages, and is adapted to domestic turkeys. *Journal of Virology* **80**(15): 7760-7764.

Spackman, E., D. A. Senne, S. Davison and D. L. Suarez (2003). Sequence analysis of recent H7 avian influenza viruses associated with three different outbreaks in commercial poultry in the United States. *Journal of Virology* **77**(24): 13399-13402.

Stadler, T. (2009). On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology* **261**(1): 58-66.

Stadler, T. (2010). Sampling-through-time in birth-death trees. *Journal of Theoretical Biology* **267**(3): 396-404.

Stadler, T., R. Kouyos, V. von Wyl, S. Yerly, J. Boni, *et al.* (2012). Estimating the Basic Reproductive Number from Viral Sequence Data. *Molecular Biology and Evolution* **29**(1): 347-357.

Stallknecht, D. E., S. M. Shane, M. T. Kearney and P. J. Zwank (1990). Persistence of Avian Influenza-Viruses in Water. *Avian Diseases* **34**(2): 406-411.

Steinhauer, D. A. (1999). Role of hemagglutinin cleavage for the pathogenicity of influenza virus. *Virology* **258**(1): 1-20.

Strimmer, K. and O. G. Pybus (2001). Exploring the Demographic History of DNA Sequences Using the Generalized Skyline Plot. *Molecular Biology and Evolution* **18**(12): 2298-2305.

Strimmer, K. and A. von Haeseler (1996). Accuracy of neighbor joining for n-taxon trees. *Systematic Biology* **45**(4): 516-523.

Strimmer, K. and A. Von Haeseler (2003). Nucleotide substitution models. *The Phylogenetic Handbook*. First edition. M. Salemi and A. M. Vandamme. Cambridge, Cambridge University Press**:** 72-97.

Strimmer, K. and A. von Haeseler (2009). Genetic distances and nucleotide substitution models. *The Phylogenetic Handbook.* Second edition. P. Lemey, M. Salemi and A. M. Vandamme. Cambridge, Cambridge University Press**:** 111-140.

Stumpf, M. P. H. and G. A. T. McVean (2003). Estimating recombination rates from population-genetic data. *Nature Reviews Genetics* **4**(12): 959-968.

Sturm-Ramirez, K. M., T. Ellis, B. Bousfield, L. Bissett, K. Dyrting*, et al.* (2004). Reemerging H5N1 influenza viruses in Hong Kong in 2002 are highly pathogenic to ducks. *Journal of Virology* **78**(9): 4892-4901.

Suarez, D. L. and D. A. Senne (2000). Sequence analysis of related low-pathogenic and highly pathogenic H5N2 avian influenza isolates from United States live bird markets and poultry farms from 1983 to 1989. *Avian Diseases* **44**(2): 356-364.

Suarez, D. L., D. A. Senne, J. Banks, I. H. Brown, S. C. Essen*, et al.* (2004). Recombination resulting in virulence shift in avian influenza outbreak, Chile. *Emerging Infectious Diseases* **10**(4): 693-699.

Suarez, D. L., E. Spackman and D. A. Senne (2003). Update on molecular epidemiology of H1, H5, and H7 influenza virus infections in poultry in North America. *Avian Diseases* **47**: 888-897.

Subbarao, E. K., W. London and B. R. Murphy (1993). A Single Amino-Acid in the Pb2-Gene of Influenza-a Virus Is a Determinant of Host Range. *Journal of Virology* **67**(4): 1761-1764.

Subbarao, K., A. Klimov, J. Katz, H. Regnery, W. Lim*, et al.* (1998). Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness. *Science* **279**(5349): 393-396.

Suchard, M. A., C. M. R. Kitchen, J. S. Sinsheimer and R. E. Weiss (2003). Hierarchical phylogenetic models for analyzing multipartite sequence data. *Systematic Biology* **52**(5): 649-664.

Suchard, M. A. and A. Rambaut (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25**(11): 1370-1376.

Sugita, S., Y. Yoshioka, S. Itamura, Y. Kanegae, K. Oguchi*, et al.* (1991). Molecular Evolution of Hemagglutinin Genes of H1N1 Swine and Human Influenza-A Viruses. *Journal of Molecular Evolution* **32**(1): 16-23.

Sugiura, N. (1978). Further Analysis of Data by Akaike's Information Criterion and Finite Corrections. *Communications in Statistics Part A - Theory and Methods* **7**(1): 13-26.

Sui, J. H., W. C. Hwang, S. Perez, G. Wei, D. Aird*, et al.* (2009). Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nature Structural & Molecular Biology* **16**(3): 265-273.

Sun, S., Q. Wang, F. Zhao, W. Chen and Z. Li (2012). Prediction of Biological Functions on Glycosylation Site Migrations in Human Influenza H1N1 Viruses. *Plos One* **7**(2): e32119.

Suzuki, Y. (2006). Natural Selection on the Influenza Virus Genome. *Molecular Biology and Evolution* **23**(10): 1902-1911.

Suzuki, Y. and T. Gojobori (1999). A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution* **16**(10): 1315-1328.

Suzuki, Y. and M. Nei (2002). Origin and evolution of influenza virus hemagglutinin genes. *Molecular Biology and Evolution* **19**(4): 501-509.

Swanson, W. J., R. Nielsen and Q. F. Yang (2003). Pervasive adaptive evolution in mammalian fertilization proteins. *Molecular Biology and Evolution* **20**(1): 18-20.

Swofford, D. L. and W. P. Maddison (1987). Reconstructing Ancestral Character States under Wagner Parsimony. *Mathematical Biosciences* **87**(2): 199-229.

Talbi, C., E. C. Holmes, P. De Benedictis, O. Faye, E. Nakoune, *et al.* (2009). Evolutionary history and dynamics of dog rabies virus in western and central Africa. *Journal of General Virology* **90**: 783-791.

Tamura, K. and M. Nei (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**: 512-526.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, *et al.* (2011). MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* **28**(10): 2731-2739.

Taubenberger, J. K. and D. M. Morens (2006). 1918 influenza: the mother of all pandemics. *Emerging Infectious Diseases* **12**(1): 15-22.

Tavaré, S. (1984). Line-of-descent and genealogical processes and their applications in population genetics models. *Theor. Popul. Biol.* **26**(2): 119-164.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures in Mathematics and Life Sciences* **17**: 57-86.

Tong, S., Y. Li, P. Rivailler, C. Conrardy, D. A. A. Castillo, *et al.* (2012). A distinct lineage of influenza A virus from bats. *Proceedings of the National Academy of Sciences*.

Vermund, S. H. and A. J. Leigh Brown (2011). The HIV epidemic: High-Income countries. *Human Immunodeficiency Virus*. F. D. Bushman, G. J. Nabel and R. Swanstrom. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press.

Vidal, N., C. Mulanga-Kabeya, N. Nzilambi, D. Robertson, W. Ilunga, *et al.* (2000). Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *Journal of Virology* **74**(22): 10498-10507.

Vijaykrishna, D., J. Bahl, S. Riley, L. Duan, J. X. Zhang, *et al.* (2008a). Evolutionary dynamics and emergence of panzootic H5N1 influenza viruses. *PLoS Pathogens* **4**(9).

Vijaykrishna, D., J. Wang, H. Chen, G. J. D. Smith, J. S. M. Peiris, *et al.* (2008b). Evolution of influenza A (H5N1) virus in Asia: evidence from systematic surveillance. *Indian Journal of Virology* **19**(1): 96-96.

Vince, M. (1996). *Softbills: care, breeding and conservation*, Hancock House.

Vincent, A. L., W. J. Ma, K. M. Lager, B. H. Janke and J. A. Richt (2008). Swine Influenza Viruses: A North American Perspective. *Advances in Virus Research, Vol 72* **72**: 127-154.

Volz, E.M. (2012). Complex population dynamics and the coalescent under neutrality. *Genetics* **190**(1):187-201

Volz, E. M., J. S. Koopman, M. J. Ward, A. J. Leigh Brown and S. D. W. Frost (2012). Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection. *PLoS Computational Biology* **8**(6): e1002552

Volz, E. M., S. L. K. Pond, M. J. Ward, A. J. L. Brown and S. D. W. Frost (2009). Phylodynamics of Infectious Disease Epidemics. *Genetics* **183**(4): 1421-1430.

Wagner, R., M. Matrosovich and H. D. Klenk (2002). Functional balance between haemagglutinin and neuraminidase in influenza virus infections. *Reviews in Medical Virology* **12**(3): 159-166.

Wagner, R., T. Wolff, A. Herwig, S. Pleschka and H. D. Klenk (2000). Interdependence of hemagglutinin glycosylation and neuraminidase as regulators of influenza virus growth: a study by reverse genetics. *Journal of Virology* **74**(14): 6316-6323.

WAHID. (2011). World Animal Health Information Database (WAHID). Retrieved 19th January, 2011, from http://web.oie.int/wahis/public.php?page=home.

Wakeley, J. (2009). *Coalescent Theory: An Introduction*. 1, Roberts and Company.

Wakeley, J. and O. Sargsyan (2009). Extensions of the Coalescent Effective Population Size. *Genetics* **181**(1): 341-345.

Wakeley, J. (2010). Natural selection and coalescent theory. *Evolution since Darwin: the first 150 years.* M. A. Bell, D. J. Futuyma, W. F. Eanes, *and* J. S. Levinton. Sinauer and Associates, Sunderland , Massachusetts: 119–149.

Wallinga, J. and M. Lipsitch (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B-Biological Sciences* **274**(1609): 599-604.

Wallinga, J. and P. Teunis (2004a). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology* **160**(6): 509-516.

Wallinga, J. and P. Teunis (2004b). Real-time tracking of infection control measures - Response. *American Journal of Epidemiology* **160**(6): 520-520.

Wang, C., K. Takeuchi, L. H. Pinto and R. A. Lamb (1993). Ion-Channel Activity of Influenza-a Virus M(2) Protein - Characterization of the Amantadine Block. *Journal of Virology* **67**(9): 5585-5594.

Webby, R. J. and R. G. Webster (2003). Are we ready for pandemic influenza? *Science* **302**(5650): 1519-1522.

Webster, R. G. (1998). Influenza: An emerging disease. *Emerging Infectious Diseases* **4**(3): 436-441.

Webster, R. G. (2004). Wet markets - a continuing source of severe acute respiratory syndrome and influenza? *Lancet* **363**(9404): 234-236.

Webster, R. G., W. J. Bean, O. T. Gorman, T. M. Chambers and Y. Kawaoka (1992). Evolution and Ecology of Influenza A Viruses. *Microbiological Reviews* **56**(1): 152-179.

Webster, R. G., S. Krauss, D. Hulse-Post and K. Sturm-Ramirez (2007). Evolution of influenza a viruses in wild birds. *Journal of Wildlife Diseases* **43**(3): S1-S6.

Webster, R. G. and W. G. Laver (1980). Determination of the Number of Nonoverlapping Antigenic Areas on Hong-Kong (H3n2) Influenza-Virus Hemagglutinin with Monoclonal-Antibodies and the Selection of Variants with Potential Epidemiological Significance. *Virology* **104**(1): 139-148.

Webster, R. G., W. G. Laver, G. M. Air and G. C. Schild (1982). Molecular Mechanisms of Variation in Influenza-Viruses. *Nature* **296**(5853): 115-121.

Webster, R. G., M. Yakhno, V. S. Hinshaw, W. J. Bean and K. G. Murti (1978). Intestinal Influenza - Replication and Characterization of Influenza-Viruses in Ducks. *Virology* **84**(2): 268-278.

Weinert, L., J. J. Welch, M. Suchard, P. Lemey, A. Rambaut, *et al.* (2012). Molecular dating of human-to-bovid host jumps by Staphylococcus aureus reveals an association with the spread of domestication. *Biology Letters*.

WHO (1980). A Revision of the System of Nomenclature for Influenza Viruses - a WHO Memorandum. *Bulletin of the World Health Organization* **58**(4): 585-591.

WHO (2005). Avian Influenza A (H5N1) Infection in Humans. *New England Journal of Medicine* **353**(13): 1374-1385.

WHO. (2010). Pandemic (H1N1) 2009 - update 101.   Retrieved 7th February, 2011, from http://www.who.int/csr/don/2010_05_21/en/index.html.

WHO. (2011a). Avian influenza cases table: 20th January 2011.   Retrieved 7th February, 2011, from http://www.who.int/csr/disease/avian_influenza/country/cases_table_2011_01_20/en/index.html.

WHO. (2011b). FAO-OIE-WHO Technical Update: Current evolution of avian influenza H5N1 viruses.   Retrieved 13th February, 2012, from http://www.who.int/influenza/human_animal_interface/tripartite_notes_H5N1.pdf.

Widjaja, L., S. L. Krauss, R. J. Webby, T. Xie and R. G. Webster (2004). Matrix gene of influenza a viruses isolated from wild aquatic birds: Ecology and emergence of influenza A viruses. *Journal of Virology* **78**(16): 8771-8779.

Wiley, D. C. and J. J. Skehel (1987). The Structure and Function of the Hemagglutinin Membrane Glycoprotein of Influenza-Virus. *Annual Review of Biochemistry* **56**: 365-394.

Wiley, D. C., I. A. Wilson and J. J. Skehel (1981). Structural Identification of the Antibody-Binding Sites of Hong-Kong Influenza Hemagglutinin and Their Involvement in Antigenic Variation. *Nature* **289**(5796): 373-378.

Wilson, D. J., D. Falush and G. McVean (2005). Germs, genomes and genealogies. *Trends in Ecology & Evolution* **20**(1): 39-45.

Wong, E. H. M., D. K. Smith, R. Rabadan, M. Peiris and L. L. M. Poon (2010). Codon usage bias and the evolution of influenza A viruses. *BMC Evolutionary Biology* **10**.

Wood, G. W., J. W. Mccauley, J. B. Bashiruddin and D. J. Alexander (1993). Deduced Amino-Acid-Sequences at the Hemagglutinin Cleavage Site of Avian Influenza-a Viruses of H5 and H7 Subtypes. *Archives of Virology* **130**(1-2): 209-217.

Woolhouse, M., M. Chase-Topping, D. Haydon, J. Friar, L. Matthews, *et al.* (2001). Epidemiology: Foot-and-mouth disease under control in the UK. *Nature* **411**(6835): 258-259.

Woolhouse, M. and A. Donaldson (2001). Managing foot-and-mouth: The science of controlling disease outbreaks. *Nature* **410**(6828): 515-516.

Worobey, M. (2001). A novel approach to detecting and measuring recombination: New insights into evolution in viruses, bacteria, and mitochondria. *Molecular Biology and Evolution* **18**(8): 1425-1434.

Worobey, M., M. Gemmel, D. E. Teuwen, T. Haselkorn, K. Kunstman, *et al.* (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**(7213): 661-U657.

Wright, Neumann and Y. Kawaoka (2007). Orthomyxoviruses. *Fields' Virology*.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**: 97-159.

Yamada, S., Y. Suzuki, T. Suzuki, M. Q. Le, C. A. Nidom, *et al.* (2006). Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors. *Nature* **444**(7117): 378-382.

Yang, C. F., M. Li, J. L. K. Mokili, J. Winter, N. M. Lubaki, *et al.* (2005a). Genetic diversification and recombination of HIV type 1 group M in Kinshasa, Democratic Republic of Congo. *Aids Research and Human Retroviruses* **21**(7): 661-666.

Yang, H., L. M. Chen, P. J. Carney, R. O. Donis and J. Stevens (2010). Structures of Receptor Complexes of a North American H7N2 Influenza Hemagglutinin with a Loop Deletion in the Receptor Binding Site. *Plos Pathogens* **6**(9).

Yang, O. O., E. S. Daar, B. D. Jamieson, A. Balamurugan, D. M. Smith*, et al.* (2005b). Human immunodeficiency virus type 1 clade B superinfection: Evidence for differential immune containment of distinct clade B strains. *Journal of Virology* **79**(2): 860-868.

Yang, Y., J. D. Sugimoto, M. E. Halloran, N. E. Basta, D. L. Chao*, et al.* (2009). The Transmissibility and Control of Pandemic Influenza A (H1N1) Virus. *Science* **326**(5953): 729-733.

Yang, Z. (1994). Estimating the Pattern of Nucleotide Substitution. *Journal of Molecular Evolution* **39**(1): 105-111.

Yang, Z., N. Goldman and A. Friday (1994). Comparison of Models for Nucleotide Substitution Used in Maximum-Likelihood Phylogenetic Estimation. *Molecular Biology and Evolution* **11**(2): 316-324.

Yang, Z. H. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**(5): 555-556.

Yang, Z. H. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**(5): 568-573.

Yang, Z. H. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**(8): 1586-1591.

Yang, Z. H. and J. P. Bielawski (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution* **15**(12): 496-503.

Yang, Z. H. and R. Nielsen (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* **46**(4): 409-418.

Yang, Z. H., R. Nielsen, N. Goldman and A. M. K. Pedersen (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**(1): 431-449.

Yang, Z. H. and W. J. Swanson (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molecular Biology and Evolution* **19**(1): 49-57.

Ypma, R. J. F., A. M. A. Bataille, A. Stegeman, G. Koch, J. Wallinga*, et al.* (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B-Biological Sciences* **279**(1728): 444-450.

Zamarin, D., M. B. Ortigoza and P. Palese (2006). Influenza A virus PB1-F2 protein contributes to viral pathogenesis in mice. *Journal of Virology* **80**(16): 7976-7983.

Zhang, J. Z., H. F. Rosenberg and M. Nei (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences of the United States of America* **95**(7): 3708-3713.

Zhu, T. F., B. T. Korber, A. J. Nahmias, E. Hooper, P. M. Sharp*, et al.* (1998). An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**(6667): 594-597.

Zhuang, J. L., A. E. Jetzt, G. L. Sun, H. Yu, G. Klarmann*, et al.* (2002). Human immunodeficiency virus type 1 recombination: Rate, fidelity, and putative hot spots. *Journal of Virology* **76**(22): 11273-11282.

Zuckerkandl, E. and L. Pauling (1965). Molecules as Documents of Evolutionary History. *Journal of Theoretical Biology* **8**(2): 357-&.

Chapter 10

Appendices

10.1 Appendix A

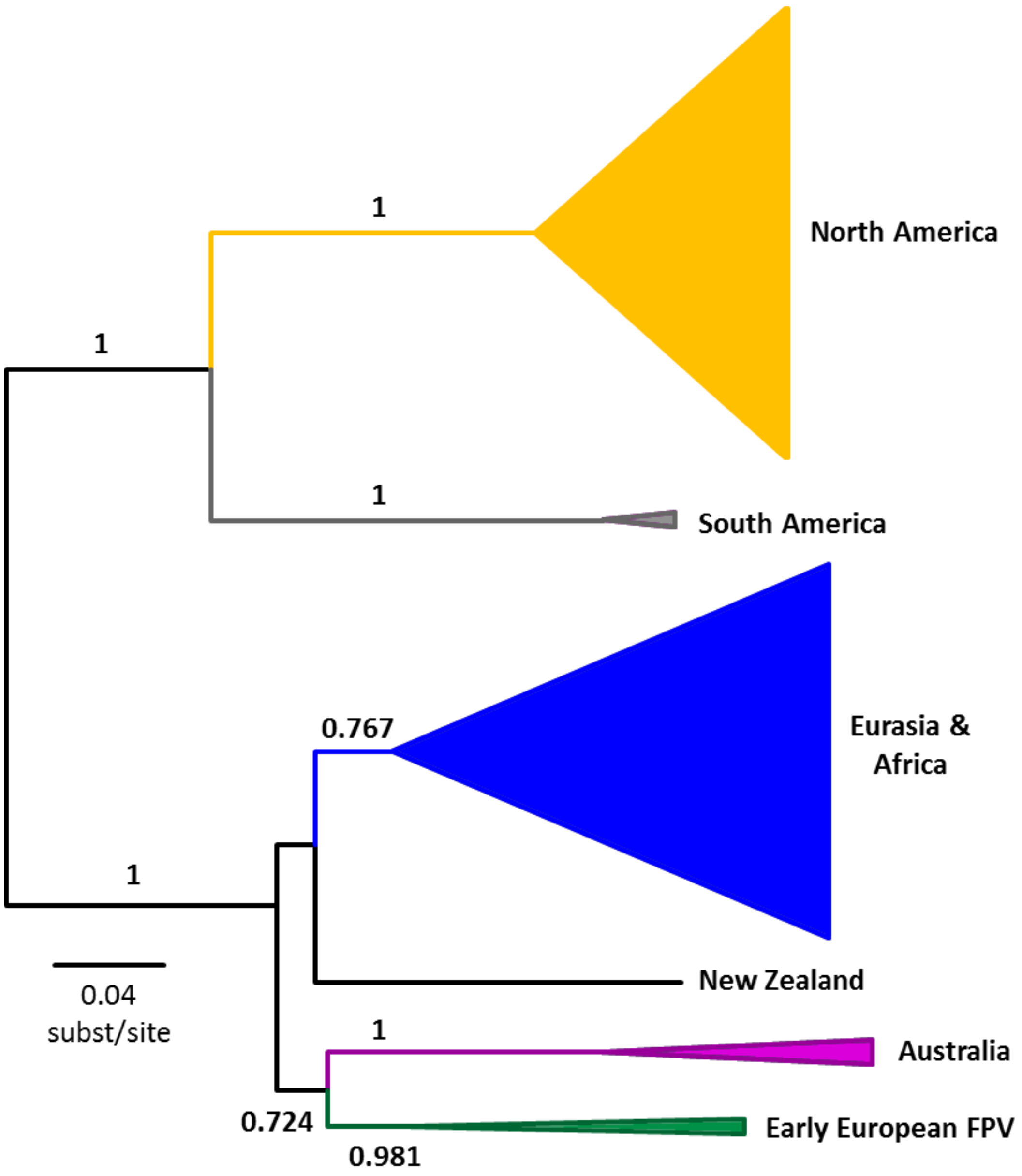


Figure A1
Neighbor-joining phylogeny of Avian H7 HA influenza sequences. The phylogeny was constructed with the MEGA software, using the neighbour-joining method with a Tamura-Nei 93 model of nucleotide substitution and gamma distributed rate heterogeneity across sites, allowing for rate heterogeneity across lineages. The tree was rooted to an H15 HA outgroup sequence (removed from figure for visualisation purposes). Clades were identified corresponding to major geographical lineages, and collapsed so that their sizes were proportional to the number of sequences at the tips of the tree in each clade. 1000 bootstrap replicates were performed and bootstrap support values are reported as the proportion of the 1000 replicates for which those sequences clustered together.

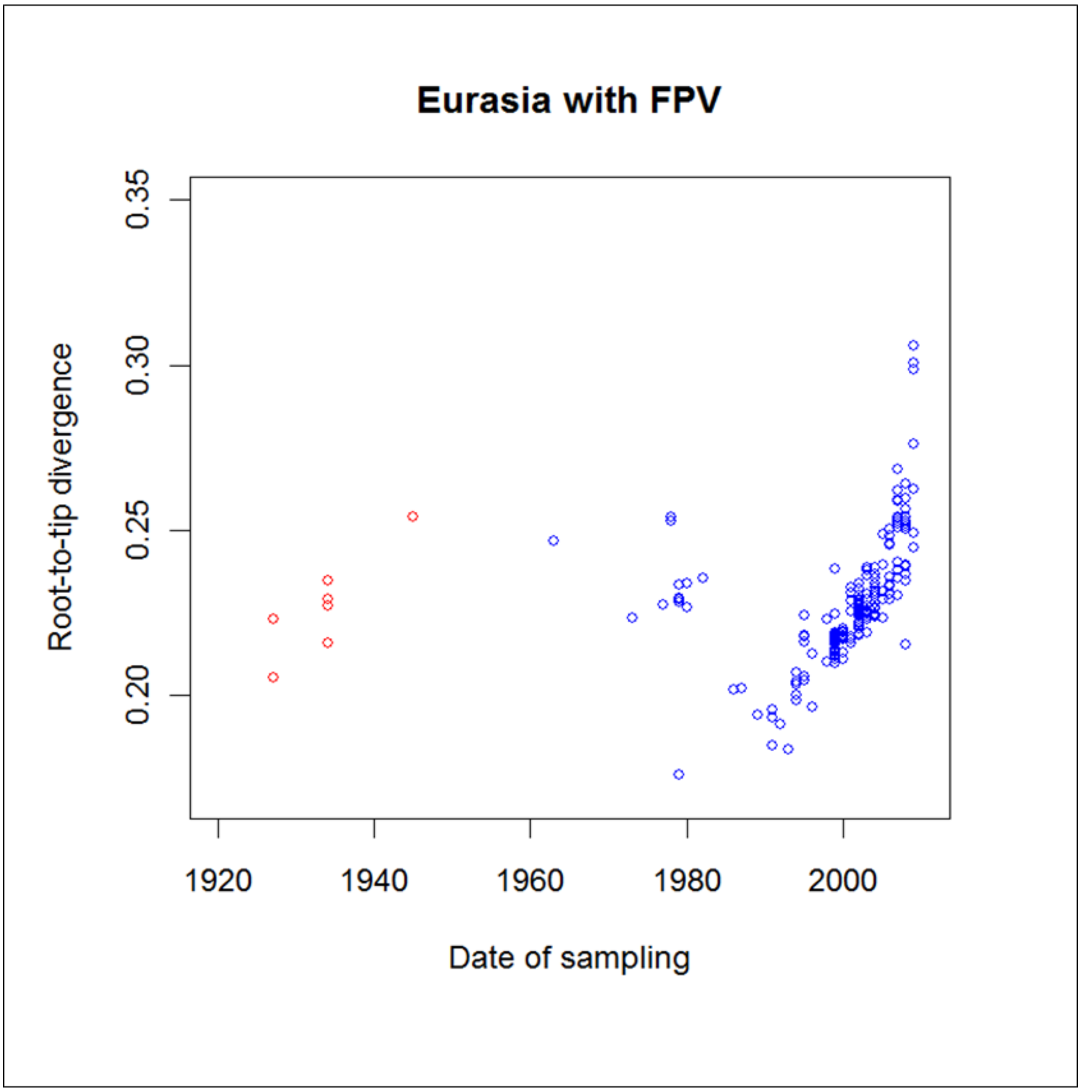


Figure A2
Root-to-tip distance plots for NJ phylogeny of Eurasian/African H7 HA influenza and early fowl plague virus sequences. For each tree tip, the distance between that tip and the root of the tree was calculated using Path-O-Gen and plotted against the year of sampling. The red circles indicate fowl plague virus sequences (HPAI isolated between 1927 and 1945), whilst the blue points represent later H7 avian influenza HA sequences.

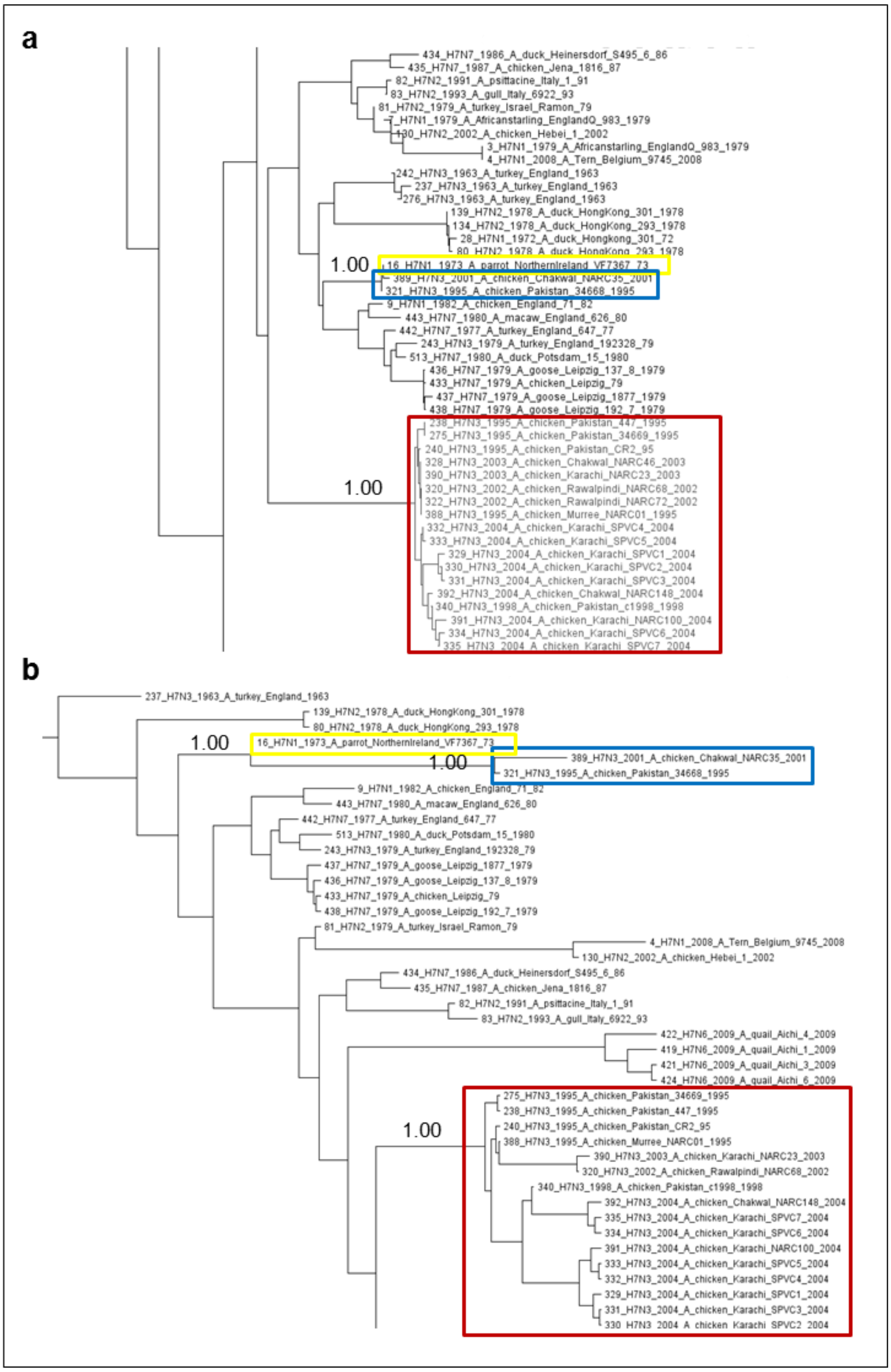

a
434_H7N7_1986_A_duck_Heinersdorf_S495_6_86
435_H7N7_1987_A_chicken_Jena_1816_87
82_H7N2_1991_A_psittacine_Italy_1_91
83_H7N2_1993_A_gull_Italy_6922_93
81_H7N2_1979_A_turkey_Israel_Ramon_79
7_H7N1_1979_A_Africanstarling_EnglandQ_983_1979
130_H7N2_2002_A_chicken_Hebei_1_2002
3_H7N1_1979_A_Africanstarling_EnglandQ_983_1979
4_H7N1_2008_A_Tern_Belgium_9745_2008
242_H7N3_1963_A_turkey_England_1963
237_H7N3_1963_A_turkey_England_1963
276_H7N3_1963_A_turkey_England_1963
139_H7N2_1978_A_duck_HongKong_301_1978
134_H7N2_1978_A_duck_HongKong_293_1978
28_H7N1_1972_A_duck_Hongkong_301_72
80_H7N2_1978_A_duck_HongKong_293_1978
1.00
16_H7N1_1973_A_parrot_NorthernIreland_VF7367_73
389_H7N3_2001_A_chicken_Chakwal_NARC35_2001
321_H7N3_1995_A_chicken_Pakistan_34668_1995
9_H7N1_1982_A_chicken_England_71_82
443_H7N7_1980_A_macaw_England_626_80
442_H7N7_1977_A_turkey_England_647_77
243_H7N3_1979_A_turkey_England_192328_79
513_H7N7_1980_A_duck_Potsdam_15_1980
436_H7N7_1979_A_goose_Leipzig_137_8_1979
433_H7N7_1979_A_chicken_Leipzig_79
437_H7N7_1979_A_goose_Leipzig_1877_1979
438_H7N7_1979_A_goose_Leipzig_192_7_1979
1.00
238_H7N3_1995_A_chicken_Pakistan_447_1995
275_H7N3_1995_A_chicken_Pakistan_34669_1995
240_H7N3_1995_A_chicken_Pakistan_CR2_95
328_H7N3_2003_A_chicken_Chakwal_NARC46_2003
390_H7N3_2003_A_chicken_Karachi_NARC23_2003
320_H7N3_2002_A_chicken_Rawalpindi_NARC68_2002
322_H7N3_2002_A_chicken_Rawalpindi_NARC72_2002
388_H7N3_1995_A_chicken_Murree_NARC01_1995
332_H7N3_2004_A_chicken_Karachi_SPVC4_2004
333_H7N3_2004_A_chicken_Karachi_SPVC5_2004
329_H7N3_2004_A_chicken_Karachi_SPVC1_2004
330_H7N3_2004_A_chicken_Karachi_SPVC2_2004
331_H7N3_2004_A_chicken_Karachi_SPVC3_2004
392_H7N3_2004_A_chicken_Chakwal_NARC148_2004
340_H7N3_1998_A_chicken_Pakistan_c1998_1998
391_H7N3_2004_A_chicken_Karachi_NARC100_2004
334_H7N3_2004_A_chicken_Karachi_SPVC6_2004
335_H7N3_2004_A_chicken_Karachi_SPVC7_2004
b
237_H7N3_1963_A_turkey_England_1963
139_H7N2_1978_A_duck_HongKong_301_1978
80_H7N2_1978_A_duck_HongKong_293_1978
1.00
16_H7N1_1973_A_parrot_NorthernIreland_VF7367_73
1.00
389_H7N3_2001_A_chicken_Chakwal_NARC35_2001
321_H7N3_1995_A_chicken_Pakistan_34668_1995
9_H7N1_1982_A_chicken_England_71_82
443_H7N7_1980_A_macaw_England_626_80
442_H7N7_1977_A_turkey_England_647_77
513_H7N7_1980_A_duck_Potsdam_15_1980
243_H7N3_1979_A_turkey_England_192328_79
437_H7N7_1979_A_goose_Leipzig_1877_1979
436_H7N7_1979_A_goose_Leipzig_137_8_1979
433_H7N7_1979_A_chicken_Leipzig_79
438_H7N7_1979_A_goose_Leipzig_192_7_1979
81_H7N2_1979_A_turkey_Israel_Ramon_79
4_H7N1_2008_A_Tern_Belgium_9745_2008
130_H7N2_2002_A_chicken_Hebei_1_2002
434_H7N7_1986_A_duck_Heinersdorf_S495_6_86
435_H7N7_1987_A_chicken_Jena_1816_87
82_H7N2_1991_A_psittacine_Italy_1_91
83_H7N2_1993_A_gull_Italy_6922_93
422_H7N6_2009_A_quail_Aichi_4_2009
419_H7N6_2009_A_quail_Aichi_1_2009
421_H7N6_2009_A_quail_Aichi_3_2009
424_H7N6_2009_A_quail_Aichi_6_2009
1.00
275_H7N3_1995_A_chicken_Pakistan_34669_1995
238_H7N3_1995_A_chicken_Pakistan_447_1995
240_H7N3_1995_A_chicken_Pakistan_CR2_95
388_H7N3_1995_A_chicken_Murree_NARC01_1995
390_H7N3_2003_A_chicken_Karachi_NARC23_2003
320_H7N3_2002_A_chicken_Rawalpindi_NARC68_2002
340_H7N3_1998_A_chicken_Pakistan_c1998_1998
392_H7N3_2004_A_chicken_Chakwal_NARC148_2004
335_H7N3_2004_A_chicken_Karachi_SPVC7_2004
334_H7N3_2004_A_chicken_Karachi_SPVC6_2004
391_H7N3_2004_A_chicken_Karachi_NARC100_2004
333_H7N3_2004_A_chicken_Karachi_SPVC5_2004
332_H7N3_2004_A_chicken_Karachi_SPVC4_2004
329_H7N3_2004_A_chicken_Karachi_SPVC1_2004
331_H7N3_2004_A_chicken_Karachi_SPVC3_2004
330_H7N3_2004_A_chicken_Karachi_SPVC2_2004

Figure A3
Position of Pakistan H7N3 avian influenza HA sequences in (a) neighbor-joining phylogeny and (b) time-scaled BEAST phylogeny. HP H7N3 Pakistan sequences are highlighted in the red boxes, LP H7N3 Pakistan sequences are highlighted in the blue boxes and the H7N1 Northern Ireland sequence with which the LP Pakistan isolates share 99% sequence homology is highlighted in the yellow box. Bootstrap support values for the neighbour-joining tree, and posterior probability values for the BEAST tree, are reported for the HP Pakistan H7N3 clade and the clade containing the LP Pakistan and Northern Ireland sequences. It is likely that the LP Pakistan isolates are laboratory contaminants, rather than representing the maintenance of separate LP and HP avian influenza virus lineages in Pakistan.

Cleavage motif	Freq.	HP/LP	Found in host species	Years	Locations	NA subtypes
PEIP----------KGR	131	LP	Anseriformes, Galliformes, Passeriformes, Struthioniformes, Pstittaciformes	1972-2009	Europe, S. Africa, Asia	N1, N2, N3, N6, N7, N8, N9
PENP----------KTR	130	LP	Charadriiformes, Galliformes	1971-2009	N. America, Central America	N1, N2, N3, N4, N5, N6, N7, N9
PEKP----------KPR	86	LP	Anseriformes, Galliformes	1998-2006	N. America	N2
PENP----------KPR	24	LP	Anseriformes, Galliformes	1995-2002	N. America	N2
PETP----------KGR	18	LP	Anseriformes, Galliformes, Psittaciformes	1989-2008	Europe, Asia, N. America*	N1, N3, N7
PEKP----------KKR	16	LP	Anseriformes, Galliformes	2002-2006	N. America	N2
PEVP----------KGR	9	LP	Anseriformes, Galliformes	1999-2002	Europe, Asia	N1, N7, N8
PEKP----------KTR	7	LP	Anseriformes, Galliformes, Charadriiformes	1995-2002	N. America, S. America	N2, N3
PEQP----------KRR	4	LP	Galliformes	2009	Asia	N6
PEIP----------KGK	3	LP	Anseriformes	2002	Europe	N7
PELP----------KGR	3	LP	Anseriformes	2009	Europe	N7, N9
PENP----------KAR	3	LP	Anseriformes	2004	N. America	N3
PEIP----------KKR	2	LP	Anseriformes	1976-2008	Australasia, Europe	N7
PEIP----------KRR	2	LP	Anseriformes, Galliformes	2005-2009	Asia, Europe	N6, N7
PEGP----------KER	1	LP	Anseriformes	2005	Australasia	N7
PEIP----------KER	1	LP	Anseriformes	1992	Asia	N7
PEIP----------RKR	1	LP	Anseriformes	2007	Australasia	N6
PEIP----------XGR	1	LP	Galliformes	2003	Europe	N3
PELP----------KRR	1	LP	Galliformes	2009	Asia	N6
PESP----------KTR	1	LP	Anseriformes	1987	N. America	N8
PGVP----------RKR	1	LP	Anseriformes	2007	Australasia	N2
PEIP------KGSRVRR	18	HP	Galliformes, Struthioniformes	1999-2000	Europe	N1

Continued on next page

PETP--------KRRKR	16	HP	Galliformes	1995-2002	Asia	N3
PEIP-------KKREKR	7	HP	Anseriformes, Galliformes, Passeriformes	1976-1993	Australasia, Asia	N7
PEIP--------KRRRR	4	HP	Galliformes	2003	Europe	N7
PEKPKTCSPLSRCRETR	4	HP	Galliformes	2002	S. America	N3
PEIP--------KKKKR	3	HP	Galliformes	1979-1992	Australasia, Europe	N3, N7
PEIP--------RKRKR	3	HP	Galliformes, Struthioniformes	1994-1997	Australasia	N3, N4
PEKPKTCSPLSRCRKTR	3	HP	Galliformes	2003	S. America	N3
PELP-------KKRRKR	3	HP	Galliformes	1927-1934	Europe	N7
PENP---KQAYRKRMTR	3	HP	Galliformes	2004-2005	N. America	N3
PEPS-------KKRKKR	3	HP	Galliformes	1934	Europe	N1
PETP--------KRRRR	3	HP	Galliformes	1963	England	N3
PEIP--------KRKKR	2	HP	Anseriformes, Galliformes	1979-2008	Europe	N7
PEIP--------RRRKR	2	HP	Galliformes	1997	Australasia	N4
PETP-------KRKRKR	2	HP	Galliformes	1995	Asia	N3
PEFS-------KKRRKR	1	HP	Galliformes	1945	Africa	N1
PEIP-------KKRKKR	1	HP	Anseriformes	1979	Europe	N7
PENP---KQAYQKRMTR	1	HP	Galliformes	2004	N. America	N3
PENP----KTTKPRPRR	1	HP	Galliformes	2007	N. America	N3
PETP------KKKKKKR	1	HP	Anseriformes	1979	Europe	N7
PEIP---------KRRR	1	??	Passeriformes	1994	N. America	N1
PENP---KQAYQKQMTR	1	??	Galliformes	2004	N. America	N3

Table A1: Cleavage site motifs from avian H7 HA influenza sequences.

All examples of cleavage site motif found amongst the avian H7 HA sequences in the NCBI influenza virus database are listed, and classified according to whether they are highly pathogenic (HP) or of low pathogenicity (LP). The number of sequences in the database which possess this motif is given, along with the range of host orders, years, geographical regions and background NA subtypes from which sequences sharing a motif have been sampled. * Corresponds to birds quarantined in N. America following transportation from Eurasia.

10.2 Appendix B

A/duck/mongolia/47/01(H7N1)
A/afristar/engq/938/79(H7N1)
A/commoniora/singapore/f89/95(H7N1)
A/fairybluebird/singapore/f92/94(H7N1)
A/africanstarling/englandq/983/79(H7N1)
A/ostrich/zimbabwe/222/96(H7N1)
A/chicken/england/71/82(H7N1)
A/fpv/egypt/45(H7N1)
A/conure/england/1234/94(H7N1)
A/parrot/england/1174/94(H7N1)
A/ostrich/southafrica/1069/91(H7N1)
A/conure/england/766/94(H7N1)
A/parakeet/netherlands/267497/94(H7N1)
A/parrot/northernireland/vf7367/73(H7N1)
A/turkey/italy/12598/99(H7N1)
A/chicken/italy/13489/99(H7N1)
A/chicken/italy/267/00(H7N1)
A/turkey/italy/3889/99(H7N1)
A/turkey/italy/4169/99(H7N1)
A/chicken/italy/4575/99(H7N1)
A/turkey/italy/4602/99(H7N1)
A/turkey/italy/4603/99(H7N1)
A/turkey/italy/3775/99(H7N1)
A/chicken/italy/445/99(H7N1)
A/chicken/italy/1067/99(H7N1)
A/mallard/alberta/34/2001(H7N1)
A/turkey/italy/4169/1999(H7N1)
A/duck/nanchang/1904/1992(H7N1)
A/chicken/italy/1285/2000(H7N1)
A/turkey/italy/3675/1999(H7N1)
A/turkey/italy/1351/2001(H7N1)

A/chicken/nj/158149/99(H7N2)
A/chicken/nj/608/02(H7N2)
A/chicken/ny/1190557/01(H7N2)
A/chicken/ny/1192567/01(H7N2)
A/chicken/ny/13878/98(H7N2)
A/chicken/ny/215868/99(H7N2)
A/chicken/ny/224094/99(H7N2)
A/chicken/ny/307493/00(H7N2)
A/chicken/pa/1490921/02(H7N2)
A/chicken/va/32/02(H7N2)
A/chicken/ny/1485812/99(H7N2)
A/guineafowl/ma/14808111/02(H7N2)
A/guineafowl/nj/132469/98(H7N2)
A/turkey/nc/11165/02(H7N2)
A/turkey/va/55/02(H7N2)
A/turkey/va/66/02(H7N2)
A/turkey/va/67/02(H7N2)
A/chicken/nj/17206/99(H7N2)
A/chicken/newjersey/20621/99(H7N2)
A/chicken/ny/3572/98(H7N2)
A/chicken/nj/15827/99(H7N2)
A/chicken/ny/13986/99(H7N2)
A/chicken/ny/341733/99(H7N2)
A/chicken/ny/147142/1999(H7N2)
A/goose/newjersey/86003/98(H7N2)
A/quail/ny/11430/99(H7N2)
A/quail/pa/20304/98(H7N2)
A/avian/ny/730636/00(H7N2)
A/chicken/hebei/1/2002(H7N2)
A/chicken/de/hobo/2004(H7N2)
A/chicken/de/viva/2004(H7N2)

A/guineafowl/italy/266184/02(H7N3)
A/mallard/italy/199/01(H7N3)
A/turkey/england/192328/79(H7N3)
A/turkey/oregon/1971(H7N3)
A/turkey/oregon/1971(H7N3)
A/widgeon/alb/284/1977(H7N3)
A/blackduck/ohio/415/2001(H7N3)
A/turkey/italy/4130/2004(H7N3)
A/chicken/pakistan/34669/1995(H7N3)
A/turkey/italy/5425/2007(H7N3)
A/turkey/england/63(H7N3)
A/chicken/chile/4977/02(H7N3)
A/turkey/chile/4418/02(H7N3)
A/turkey/italy/2987/2003(H7N3)
A/turkey/italy/3337/2004(H7N3)
A/chicken/queensland/1994(H7N3)
A/chicken/newyork/1227311/1999(H7N3)
A/mallard/italy/33/01(H7N3)
A/chicken/britishcolumbia/04(H7N3)
A/greenwingedteal/alb/228/1985(H7N3)
A/mallardduck/alberta/435/1985(H7N3)
A/turkey/italy/2685/2003(H7N3)
A/turkey/italy/3477/2004(H7N3)
A/turkey/oregon/1971(H7N3)
A/turkey/tennessee/1/79(H7N3)
A/turkey/oregon/1971(H7N3)
A/ruddyturnstone/nj/65/1985(H7N3)
A/chicken/england/4266/2006(H7N3)
A/turkey/italy/251/2003(H7N3)
A/chicken/chile/4322/02(H7N3)
A/chicken/britishcolumbia/gsc_human_b/04(H7N3)

Continued on next page

A/turkey/italy/2984/2000(H7N1)
A/turkey/italy/4426/2000(H7N1)
A/chicken/italy/322/2001(H7N1)
A/duck/italy/551/2000(H7N1)
A/guineafowl/italy/155/2000(H7N1)
A/quail/italy/396/2000(H7N1)
A/chicken/italy/1082/1999(H7N1)
A/turkey/italy/977/1999(H7N1)
A/chicken/italy/2335/2000(H7N1)
A/turkey/italy/1084/2000(H7N1)
A/turkey/italy/4708/1999(H7N1)
A/turkey/italy/4295/1999(H7N1)
A/turkey/italy/3488/1999(H7N1)
A/pekinduck/italy/1848/2000(H7N1)
A/turkey/italy/4644/1999(H7N1)
A/quail/italy/4992/1999(H7N1)
A/turkey/italy/4294/1999(H7N1)
A/turkey/italy/4617/1999(H7N1)
A/turkey/italy/4301/1999(H7N1)
A/turkey/italy/3489/1999(H7N1)
A/turkey/italy/3560/1999(H7N1)
A/turkey/italy/2715/1999(H7N1)
A/turkey/italy/2732/1999(H7N1)
A/turkey/italy/1265/1999(H7N1)
A/duck/hongkong/301/72(H7N1)
A/ostrich/italy/2332/00(H7N1)
A/ostrich/italy/984/00(H7N1)
A/rhea/northcarolina/39482/1993(H7N1)
A/mallard/italy/250/02(H7N1)
A/fpv/rostock/1934(H7N1)
A/ts1/1/a/fpv/rostock/1934(H7N1)
A/duck/hongkong/301/1978(H7N2)
A/chicken/newyork/131425/94(H7N2)
A/chicken/md/minhma/2004(H7N2)
A/dk/hongkong/293/1978(H7N2)
A/chicken/pennsylvania/143586/2002(H7N2)
A/quail/italy/4610/2003(H7N2)
A/chukar/newyork/116531/2005(H7N2)
A/chicken/newyork/16330/2005(H7N2)
A/chicken/newyork/212112/2005(H7N2)
A/duck/newyork/212116/2005(H7N2)
A/chicken/newyork/212111/2005(H7N2)
A/chukar/newyork/212117/2005(H7N2)
A/duck/newyork/1436465/2005(H7N2)
A/chicken/newyork/31815/2006(H7N2)
A/guineafowl/newyork/83911/2006(H7N2)
A/chicken/newyork/83912/2006(H7N2)
A/guineafowl/newyork/195014/2006(H7N2)
A/chicken/newyork/290474/2006(H7N2)
A/chicken/newyork/163264/2005(H7N2)
A/chicken/ny/31815/06(H7N2)
A/guineafowl/ny/464918/2006(H7N2)
A/chicken/wales/1306/2007(H7N2)
A/duck/hongkong/293/78(H7N2)
A/chicken/newyork/147149/1999(H7N3)
A/turkey/england/1963(H7N3)
A/turkey/italy/8912/2002(H7N3)
A/chicken/pakistan/447/95(H7N3)
A/chicken/victoria/1/92(H7N3)
A/chicken/pakistan/cr2/95(H7N3)
A/chicken/queensland/667/95(H7N3)
A/chicken/chile/4957/02(H7N3)
A/chicken/chile/4968/02(H7N3)
A/mallard/netherlands/12/2000(H7N3)
A/turkey/minnesota/1200/1980(H7N3)
A/mallard/ohio/322/1998(H7N3)
A/mallard/alberta/24/01(H7N3)
A/mallard/alberta/24/01(H7N3)
A/pheasant/minnesota/917/1980(H7N3)
A/chicken/victoria/224/1992(H7N3)
A/chicken/italy/270638/02(H7N3)
A/duck/taiwan/33/1993(H7N7)
A/duck/taiwan/ya103/1993(H7N7)
A/turkey/ireland/pv74/1995(H7N7)
A/nonpsittacine/englandq/1985/89(H7N7)
A/turkey/northernireland/vf1545c5/98(H7N7)
A/turkey/england/647/77(H7N7)
A/macaw/england/626/80(H7N7)
A/ostrich/southafrica/m320/96(H7N7)
A/chicken/germany/r28/03(H7N7)
A/netherlands/127/03(H7N7)
A/chicken/netherlands/1/03(H7N7)
A/netherlands/219/03(H7N7)
A/mallard/sweden/56/02(H7N7)
A/mallard/sweden/82/02(H7N7)
A/mallard/sweden/85/02(H7N7)
A/mallard/sweden/87/02(H7N7)
A/mallard/sweden/92/02(H7N7)
A/mallard/sweden/93/02(H7N7)
A/mallard/sweden/94/02(H7N7)
A/mallard/sweden/100/02(H7N7)
A/mallard/sweden/102/02(H7N7)
A/mallard/sweden/103/02(H7N7)
A/mallard/sweden/104/02(H7N7)
A/mallard/sweden/105/02(H7N7)
A/mallard/sweden/106/02(H7N7)
A/mallard/sweden/107/02(H7N7)
A/ruddyturnstone/de/2378/1988(H7N7)
A/redknot/nj/325/1989(H7N7)

Continued on next page

A/turkey/newyork/44505/94(H7N2)
A/chicken/newyork/138337/95(H7N2)
A/chicken/newyork/80302/96(H7N2)
A/chicken/pennsylvania/117671/97(H7N2)
A/chicken/newyork/67773/97(H7N2)
A/turkey/pennsylvania/7975/97(H7N2)
A/chicken/pennsylvania/135521/98(H7N2)
A/quail/newyork/1398951/98(H7N2)
A/turkey/israel/ramon/79(H7N2)
A/psittacine/italy/1/91(H7N2)
A/gull/italy/6922/93(H7N2)
A/avian/ny/1183531/2001(H7N2)
A/avian/ny/7041112/00(H7N2)
A/avian/ny/742112/00(H7N2)
A/avian/ny/762473/00(H7N2)
A/avian/ny/817465/00(H7N2)
A/avian/ny/77296/00(H7N2)
A/chicken/fl/903484/01(H7N2)
A/chicken/nj/1188785/01(H7N2)
A/chicken/nj/1503837/02(H7N2)
A/chicken/nj/15124418/02(H7N2)

A/bluewingedteal/ohio/658/2004(H7N3)
A/turkey/england/63(H7N3)
A/turkey/italy/3620/2003(H7N3)
A/turkey/italy/1010/2003(H7N3)
A/mallard/delaware/418/2005(H7N3)
A/turkey/utah/2472110/1995(H7N3)
A/chicken/britishcolumbia/cn7/2004(H7N3)
A/mallardduck/alb/279/1977(H7N3)
A/shorebird/delaware/22/06(H7N3)
A/laughinggull/delaware/42/06(H7N3)
A/turkey/italy/9742/2002(H7N3)
A/turkey/italy/3829/2004(H7N3)
A/gsc_chicken_b/britishcolumbia/04(H7N3)
A/chicken/chile/1842404322/2002(H7N3)
A/chicken/england/4054/2006(H7N3)
A/turkey/italy/4479/2004(H7N3)
A/turkey/italy/4608/2003(H7N3)
A/turkey/italy/2043/2003(H7N3)
A/chicken/chile/176822/02(H7N3)
A/cinnamonteal/bolivia/4537/2001(H7N3)
A/turkey/italy/8000/2002(H7N3)

A/fowl/dobson/1927(H7N7)
A/starling/victoria/1985(H7N7)
A/chicken/victoria/1976(H7N7)
A/chicken/victoria/1/1985(H7N7)
A/chicken/netherlands/03010132/03(H7N7)
A/mallard/italy/299/05(H7N7)
A/duck/jiangxi/1742/03(H7N7)
A/fpv/weybridge(H7N7)
A/goose/leipzig/137/8/1979(H7N7)
A/goose/leipzig/187/7/1979(H7N7)
A/goose/leipzig/192/7/1979(H7N7)
A/chicken/victoria/1/1985(H7N7)
A/starling/victoria/1/1985(H7N7)
A/chicken/leipzig/79(H7N7)
A/duck/heinersdorf/s495/6/86(H7N7)
A/chicken/jena/1816/87(H7N7)
A/fpv/dutch/27(H7N7)
A/chicken/ireland/1733/89(H7N7)
A/chicken/victoria/75(H7N7)

Table B1

Names of H7 avian HA sequences analysed. All full-length avian influenza H7 HA sequences were downloaded from the NCBI database in April 2008, and are listed here by their standard sequence names. After identical nucleotide sequences were excluded, all remaining sequences were analysed in MrBayes, although further sequences (such as the old Eurasian Fowl Plague Virus sequences) were later excluded from the mutational mapping analysis.

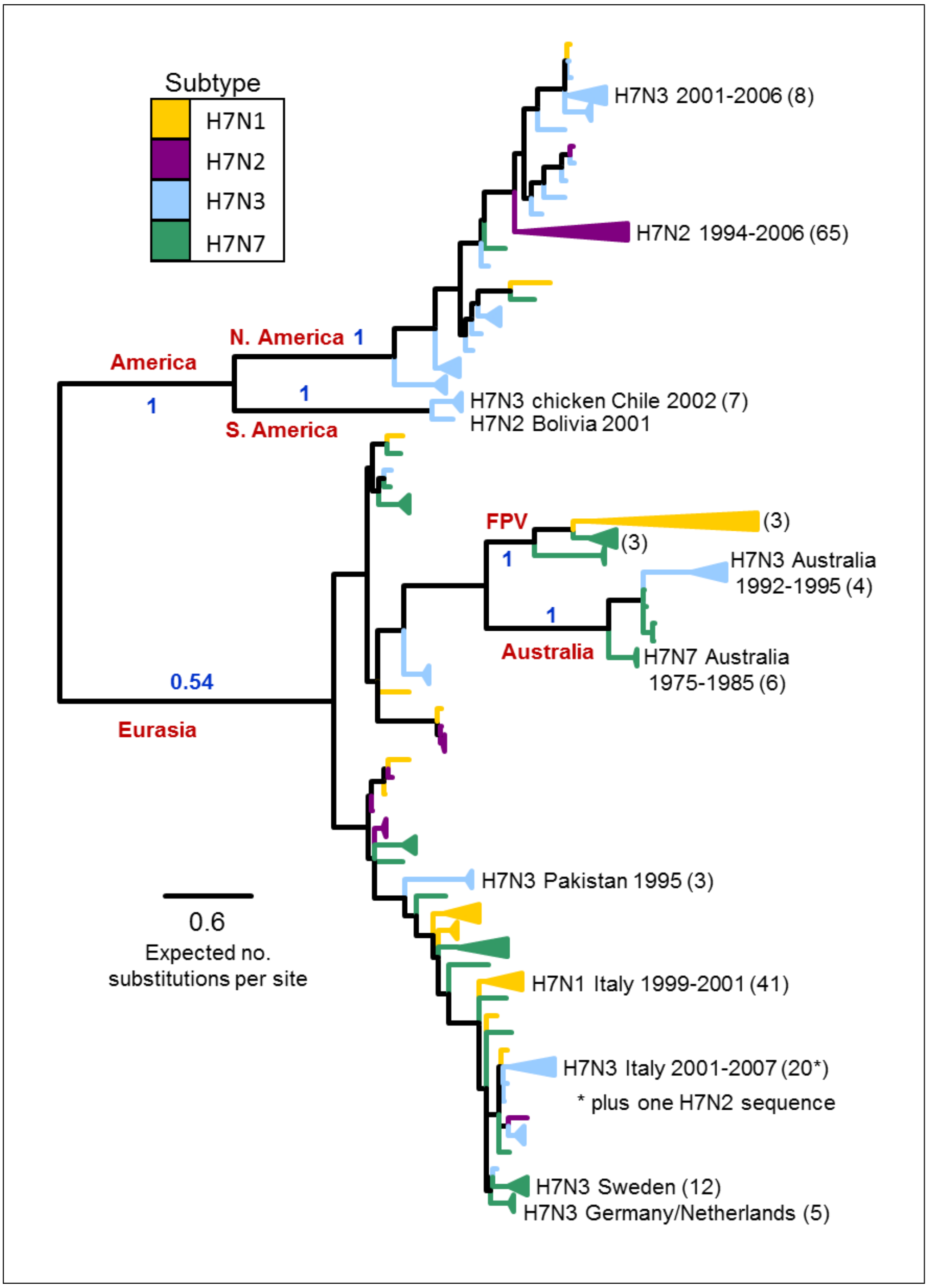


Figure B1

H7 HA1 MrBayes consensus phylogeny. The tree was inferred under the GTR + Γ model of DNA substitution, with 6 rate categories, and constructed from 1000 post-burnin MCMC phylogeny samples from MrBayes. Major geographical lineages are labelled in red and posterior probabilities of clades are labelled in blue. An H15 sequence was used as an outgroup in the phylogenetic analysis, but removed in this figure for the purpose of presentation. Lineages are coloured by the background NA subtype of the virus at the tips of the tree, and clades of sequences of the same subtype have been collapsed for the purpose of presentation (numbers of sequences in collapsed clades are given in brackets). Note: FPV = ‘fowl plague virus’, a term used to describe H7 avian influenza viruses isolated in the 1920s-1940s.

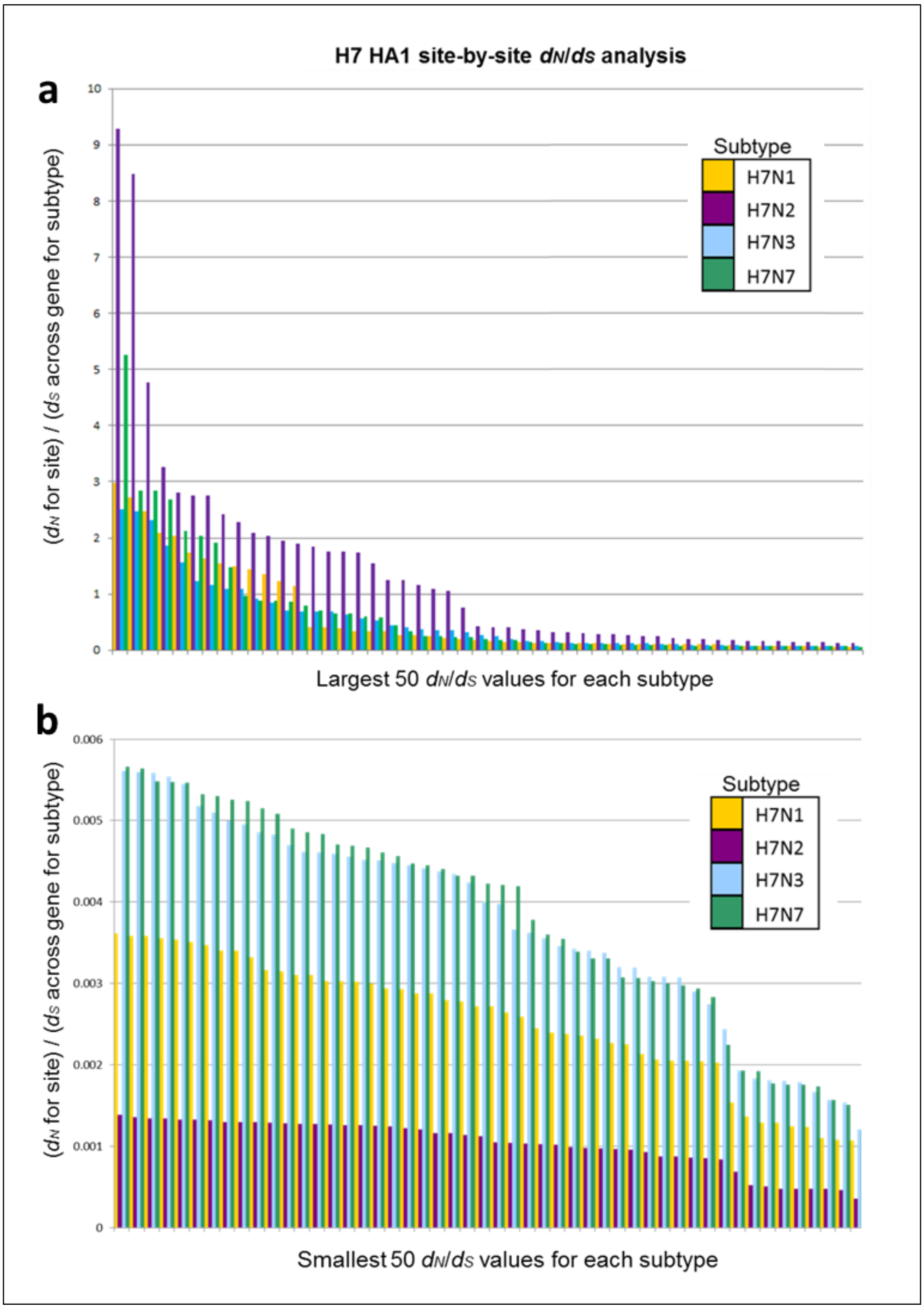


Figure B2

Site-by-site d_N/d_S values across the avian influenza H7 HA1, ranked by size. For each NA background subtype, the d_N value for each site was divided by the average d_S across all sites for that subtype. The site-by-site d_N/d_S values were ranked by size: (a) the largest 50 values were plotted for each subtype and (b) the smallest 50 values were plotted for each subtype. For all of the largest 50 d_N/d_S values, d_N/d_S on the N2 NA background was larger than the values of the same rank on the N1, N3 or N7 NA backgrounds. For all of the smallest 50 d_N/d_S values for the H7 HA1 sites, the value of d_N/d_S on the N2 NA background was smaller than the values of the same rank on the N1, N3 or N7 NA backgrounds.

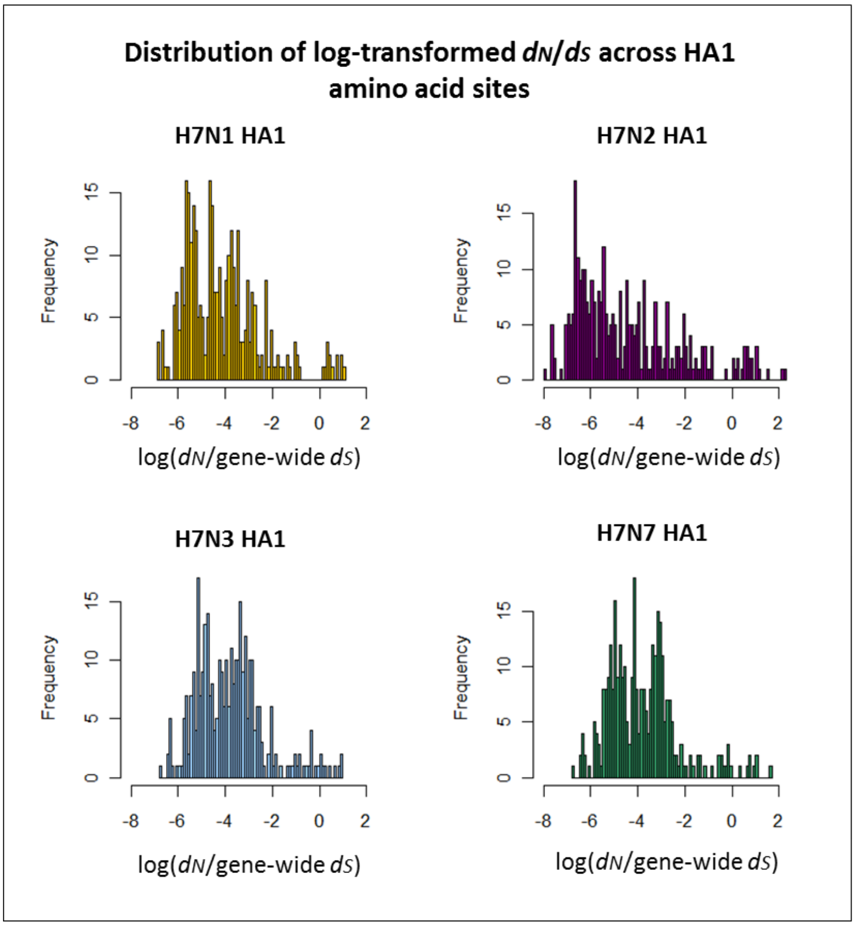


Figure B3

Histograms showing frequency of different log(d_N/gene-wide d_S) values across the H7 HA1 alignment for H7N1, H7N2, H7N3 and H7N7 lineages. Sites with log(d_N/d_S) > 0 correspond to d_N/d_S > 1, and sites with log(d_N/d_S) < 0 correspond to d_N/d_S < 1. H7N2 was the only subtype for which log(d_N/d_S) values less than -7, or greater than 2, was observed.

Subtype	Mean	Lower 90% HPD limit	Upper 90% HPD limit
H7N1	0.107	0.060	0.154
H7N2	0.189	0.108	0.253
H7N3	0.092	0.057	0.122
H7N7	0.105	0.057	0.144

Table B2
Comparing d_N/d_S for H7 HA1 (not including signal peptide region) avian influenza on different NA backgrounds. Means and 90% HPD limits of the posterior distributions for d_N/d_S were averaged across sites in the alignment corresponding to the HA1 coding region only (not including the 17 amino acid signal peptide region). The same ordering of d_N/d_S values between different NA backgrounds was observed whether the signal peptide region was included (Table 4.2) or excluded.

Comparison	d_N/d_S	d_N	d_S
H7N1-H7N2	0.058538 0.941462	0.10496 0.89504	0.565277 0.434723
H7N1-H7N3	0.646575 0.353425	0.619929 0.380071	0.492532 0.507468
H7N1-H7N7	0.514085 0.485915	0.484746 0.515254	0.470188 0.529812
H7N2-H7N3	0.982999 0.017001	0.953252 0.046748	0.424874 0.575126
H7N2-H7N7	0.950872 0.049128	0.890585 0.109415	0.404126 0.595874
H7N3-H7N7	0.362081 0.637919	0.365771 0.634229	0.478662 0.521338

Table B3
Comparing posterior distributions of evolutionary rates for avian influenza HA1 (not including signal peptide region) across different background NA subtypes. The proportion of randomised pairings of posterior rate estimates for which the value for the first subtype in the comparison, minus the value for the second subtype in the comparison, was greater than 0 (top value in cell) and less than 0 (bottom value in cell) was reported (cells highlighted yellow indicate a split at least as extreme as 5%, and cells highlighted orange indicate a split of between 5% and 10%). Here, only sites in the HA1 coding region were included, whereas in Table 4.3 the results were averaged over the HA1 coding region and the 17 amino acid signal peptide region. All comparisons involving the N2 NA background indicated that d_N/d_S was higher on the N2 background than on the N1, N3 or N7 background, as was the case when the signal peptide region was included. However, the difference in the locations of the distributions for d_N/d_S and d_N was slightly less pronounced when the signal peptide region was excluded.

10.3 Appendix C

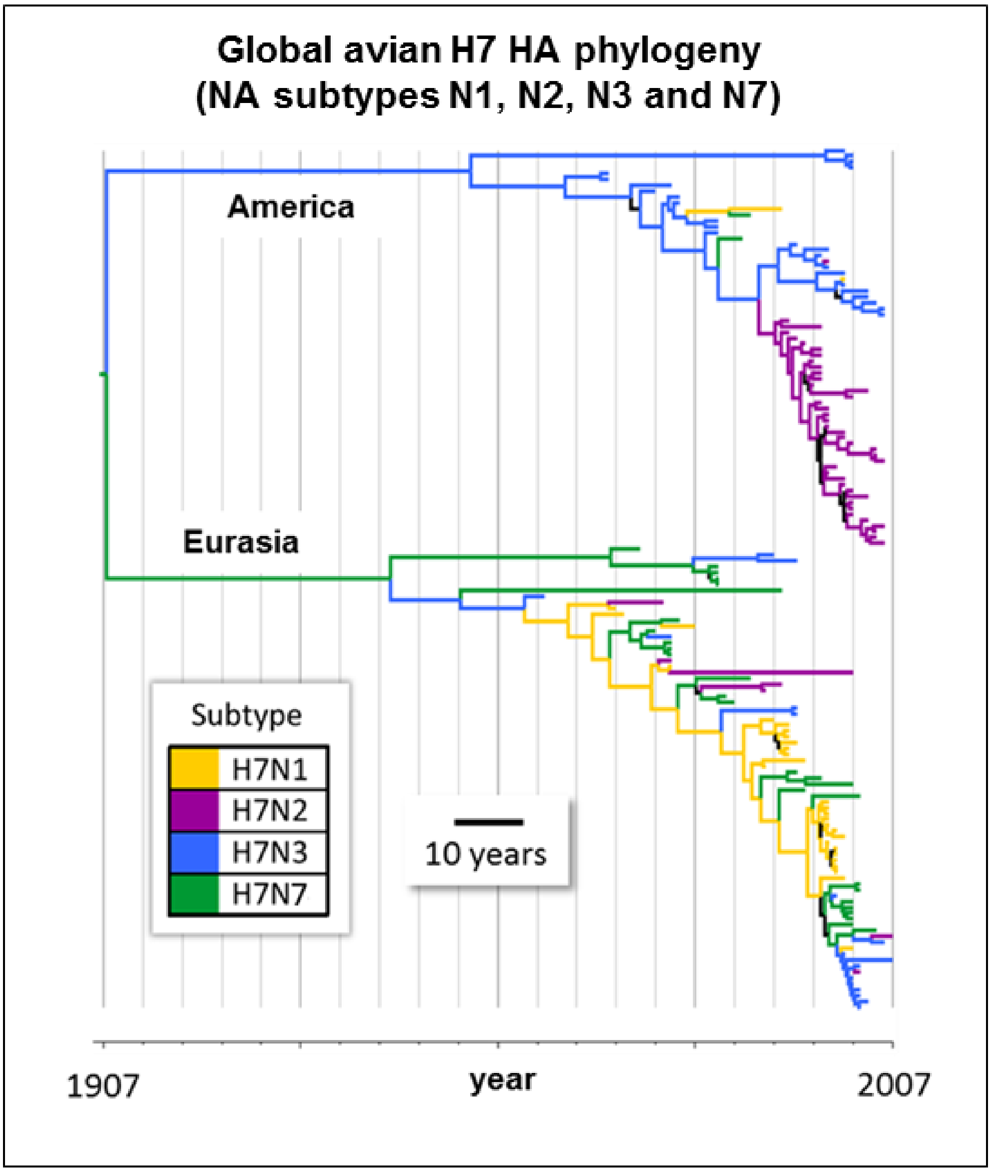


Figure C1
Discrete trait mapping of background viral NA subtypes upon global avian H7 HA sequences (N1, N2, N3 and N7 NA only). Analysis was performed upon the dataset described in Chapter 4 (with subsampling within large monophyletic clades of the same subtype). Discrete trait mapping was carried out in BEAST and the branches of the maximum clade credibility tree were coloured according to the inferred viral NA subtype at the parental node of the branch. The long divergence between the American and Eurasian lineages, and the different distribution of subtypes, suggests that discrete trait transition should be analysed separately in these regions.

H7N1_1995_A/commoniora/Singapore/F89/95_wild_LP
H7N1_1994_A/fairybluebird/Singapore/F92/94_wild_LP
H7N1_1996_A/ostrich/Zimbabwe/222/96_dom_LP
H7N1_1994_A/conure/England/1234/94_dom_LP
H7N1_1994_A/parrot/England/1174/94_dom_LP
H7N1_1991_A/ostrich/SouthAfrica/1069/91_dom_LP
H7N1_1994_A/conure/England/766/94_dom_LP
H7N1_1994_A/parakeet/Netherlands/267497/94_dom_LP
H7N1_1999_A/turkey/Italy/12598/99_dom_HP
H7N1_1999_A/chicken/Italy/13489/99_dom_HP
H7N1_2000_A/chicken/Italy/267/0_dom_HP
H7N1_1999_A/turkey/Italy/3889/99_dom_LP
H7N1_1999_A/turkey/Italy/4073/99_dom_LP
H7N1_1999_A/turkey/Italy/4169/1999_dom_LP
H7N1_1999_A/chicken/Italy/4575/99_dom_LP
H7N1_1999_A/turkey/Italy/4602/99_dom_LP
H7N1_1999_A/turkey/Italy/4603/1999_dom_LP
H7N1_1999_A/chicken/Italy/445/99_dom_HP
H7N1_1999_A/chicken/Italy/1067/1999_dom_LP
H7N1_2000_A/chicken/Italy/1285/2000_dom_HP
H7N1_2000_A/ostrich/Italy/2332/0_dom_HP
H7N1_1999_A/turkey/Italy/4580/1999_dom_HP
H7N1_1999_A/turkey/Italy/3675/1999_dom_LP
H7N1_2001_A/turkey/Italy/1351/2001_dom_LP
H7N1_2000_A/turkey/Italy/2984/2000_dom_HP
H7N1_2001_A/chicken/Italy/322/2001_dom_LP
H7N1_2000_A/duck/Italy/551/2000_dom_HP
H7N1_2000_A/chicken/Italy/2335/2000_dom_HP
H7N1_2000_A/turkey/Italy/1084/2000_dom_HP
H7N1_1999_A/turkey/Italy/4708/1999_dom_HP
H7N1_1999_A/turkey/Italy/4482/1999_dom_LP
H7N1_1999_A/turkey/Italy/4295/1999_dom_LP
H7N1_1999_A/turkey/Italy/3185/1999_dom_LP
H7N1_2000_A/pekinduck/Italy/1848/2000_dom_HP
H7N1_1999_A/quail/Italy/4992/1999_dom_HP
H7N1_1999_A/turkey/Italy/4294/1999_dom_LP
H7N1_1999_A/turkey/Italy/4301/1999_dom_LP
H7N1_1999_A/turkey/Italy/3489/1999_dom_LP
H7N1_1999_A/turkey/Italy/3560/1999_dom_LP
H7N1_1999_A/turkey/Italy/2715/1999_dom_LP
H7N1_1999_A/turkey/Italy/2732/1999_dom_LP
H7N1_1999_A/turkey/Italy/1265/1999_dom_LP
H7N1_2002_A/mallard/Italy/250/2_dom_LP
H7N1_2003_A/duck/Hokkaido/143/2003_wild_LP
H7N1_2008_A/duck/Denmark/531478/2008_dom_LP
H7N1_2007_A/mallard/Netherlands/22/2007_wild_LP
H7N1_1991_A/ostrich/SouthAfrica/1991/dom_LP_
H7N1_1999_A/turkey/Italy/3283/1999_dom_LP
H7N1_1999_A/turkey/Italy/4426/1999_dom_LP
H7N1_2000_A/turkey/Italy/2379/2000_dom_LP
H7N2_2003_A/quail/Italy/4610/2003_dom_LP
H7N2_2007_A/chicken/Wales/1306/2007_dom_LP
H7N2_2003_A/wildbirdfeces/Korea/HDR16/2003_wild_LP
H7N2_2004_A/wildbirdfeces/Nakdonggang/214/2004_wild_LP

Continued on next page

H7N1_2001_A/duck/Mongolia/47/2001_wild_LP
H7N1_2000_A/ostrich/Italy/1038/2000_dom_HP
H7N1_2000_A/guineafowl/Italy/155/2000_dom_HP
H7N1_2000_A/quail/Italy/396/2000_dom_HP
H7N1_1999_A/chicken/Italy/1082/99_dom_LP
H7N1_1999_A/turkey/Italy/977/1999_dom_LP
H7N3_2001_A/mallard/Italy/43/2001_wild_LP
H7N3_1995_A/chicken/Pakistan/34669/1995_dom_HP
H7N3_2006_A/chicken/England/4266/2006_dom_LP
H7N3_2006_A/chicken/England/4054/2006_dom_LP
H7N3_2004_A/turkey/Italy/4479/2004_dom_LP
H7N3_2003_A/turkey/Italy/251/2003_dom_LP
H7N3_2004_A/turkey/Italy/3807/2004_dom_LP
H7N3_2002_A/turkey/Italy/8912/2002_dom_LP
H7N3_2003_A/turkey/Italy/3620/2003_dom_LP
H7N3_2003_A/turkey/Italy/1010/2003_dom_LP
H7N3_2003_A/turkey/Italy/4608/2003_dom_LP
H7N3_2003_A/turkey/Italy/2987/2003_dom_LP
H7N3_2004_A/turkey/Italy/3337/2004_dom_LP
H7N3_2003_A/turkey/Italy/2043/2003_dom_LP
H7N3_2002_A/turkey/Italy/8000/2002_dom_LP
H7N3_2002_A/Guineafowl/Italy/266184/2_dom_LP
H7N3_2001_A/mallard/Italy/199/1_wild_LP
H7N3_2002_A/turkey/Italy/9742/2002_dom_LP
H7N3_2003_A/turkey/Italy/2685/2003_dom_LP
H7N3_2004_A/turkey/Italy/3829/2004_dom_LP
H7N3_2004_A/turkey/Italy/4130/2004_dom_LP
H7N2_2003_A/wildbirdfeces/Hadoree/8/2003_wild_LP
H7N2_2006_A/mallard/Netherlands/29/2006_wild_LP
H7N3_1995_A/chicken/Pakistan/447/1995_dom_HP
H7N3_1995_A/chicken/Pakistan/CR2/95_dom_HP
H7N3_2000_A/mallard/Netherlands/12/2000_wild_LP
H7N3_2002_A/turkey/Italy/214845/2002_dom_LP
H7N3_2008_A/Northernshoveler/Seongdong/175/2008_wild_LP
H7N3_2006_A/wildbirdfeces/Shihwa/21/2006_wild_LP
H7N3_1995_A/chicken/Murree/NARC01/1995_dom_HP
H7N3_2003_A/chicken/Karachi/NARC23/2003_dom_HP
H7N3_2004_A/chicken/Karachi/NARC100/2004_dom_HP
H7N3_2004_A/chicken/Chakwal/NARC148/2004_dom_HP
H7N3_2006_A/tuftedduck/PT/13771/2006_wild_LP
H7N7_1995_A/turkey/Ireland/PV74/1995_dom_LP
H7N7_1998_A/turkey/NorthernIreland/VF1545C5/98_dom_LP
H7N7_1996_A/ostrich/SouthAfrica/M320/96_dom_LP
H7N7_2003_A/chicken/Germany/R28/3_dom_HP
H7N7_2002_A/mallard/Sweden/56/2002_wild_LP
H7N7_2002_A/mallard/Sweden/82/2_wild_LP
H7N7_2002_A/mallard/Sweden/85/2002_wild_LP
H7N7_2002_A/mallard/Sweden/87/2_wild_LP
H7N7_2002_A/mallard/Sweden/92/2_wild_LP
H7N7_2002_A/mallard/Sweden/93/2_wild_LP
H7N7_2002_A/mallard/Sweden/94/2_wild_LP
H7N7_2002_A/mallard/Sweden/100/2_wild_LP
H7N7_2002_A/mallard/Sweden/102/2_wild_LP
H7N7_2002_A/mallard/Sweden/103/2_wild_LP

Continued on next page

H7N3_2007_A/turkey/Italy/5425/2007_dom_LP
H7N3_2002_A/turkey/Italy/9739/2002_dom_LP
H7N3_2003_A/chicken/Italy/682/2003_dom_LP
H7N3_2002_A/chicken/Rawalpindi/NARC68/2002_dom_HP
H7N3_2004_A/chicken/Karachi/SPVC1/2004_dom_HP
H7N3_2004_A/chicken/Karachi/SPVC2/2004_dom_HP
H7N3_2004_A/chicken/Karachi/SPVC3/2004_dom_HP
H7N3_2004_A/chicken/Karachi/SPVC4/2004_dom_HP
H7N3_2004_A/chicken/Karachi/SPVC5/2004_dom_HP
H7N3_2004_A/chicken/Karachi/SPVC6/2004_dom_HP
H7N3_2004_A/chicken/Karachi/SPVC7/2004_dom_HP
H7N3_1998_A/chicken/Pakistan/c1998/1998_dom_HP
H7N7_2007_A/mallard/Korea/GH170/2007_wild_LP
H7N7_2007_A/mallard/Korea/GG2/2007_wild_LP
H7N7_2007_A/magpie/Korea/YJD174/2007_wild_LP
H7N7_2008_A/mallard/Sweden/100993/2008_wild_LP
H7N7_2003_A/mallard/Sweden/S90735/2003_wild_LP
H7N7_2005_A/mallard/Sweden/S90597/2005_wild_LP
H7N7_2007_A/mallard/Geumgang/1/2007_wild_LP
H7N7_2007_A/muteswan/Hungary/5973/2007_wild_LP
H7N7_2008_A/northernpintail/Miyagi/674/2008_wild_LP
H7N7_2008_A/northernpintail/Aomori/1001/2008_wild_LP
H7N7_2008_A/northernpintail/Akita/1368/2008_wild_LP
H7N7_2008_A/northernpintail/Akita/1369/2008_wild_LP
H7N7_2008_A/northernpintail/Akita/1370/2008_wild_LP
H7N7_2009_A/swan/Slovenia/53/2009_wild_LP
H7N7_2002_A/mallard/Sweden/104/2_wild_LP
H7N7_2002_A/mallard/Sweden/105/2002_wild_LP
H7N7_2002_A/mallard/Sweden/106/2_wild_LP
H7N7_2002_A/mallard/Sweden/107/2_wild_LP
H7N7_2003_A/chicken/Netherlands/3010132/3_dom_HP
H7N7_2003_A/duck/Jiangxi/1742/3_dom_LP
H7N7_2005_A/mallard/Italy/299/5_wild_LP
H7N7_2003_A/chicken/Netherlands/2586/2003_dom_HP
H7N7_1999_A/duck/Taiwan/4201/99_wild_LP
H7N7_2008_A/chicken/England/115811406/2008_dom_HP
H7N7_2006_A/wildbirdfeces/Korea/HDR22/2006_wild_LP
H7N7_2005_A/wildbirdfeces/Korea/ESD07/2005_wild_LP
H7N7_2007_A/duck/Shiga/B149/2007_wild_LP
H7N7_2007_A/duck/Tsukuba/30/2007_wild_LP
H7N7_2007_A/duck/Tsukuba/700/2007_wild_LP
H7N7_2008_A/duck/Tsukuba/922/2008_wild_LP
H7N7_2009_A/duck/Chiba/20/2009_wild_LP
H7N7_2005_A/mallard/Netherlands/9/2005_wild_LP
H7N8_1999_A/swan/Shimane/42/1999_wild_LP
H7N8_2006_A/mallard/Netherlands/33/2006_wild_LP
H7N8_2008_A/garganey/Crimea/2027/2008_wild_LP
H7N9_2002_A/mallard/Sweden/91/2_wild_LP
H7N9_2008_A/duck/Mongolia/119/2008_wild_LP
H7N9_2008_A/Anascrecca/Spain/1460/2008_wild_LP
H7N9_2009_A/goose/CzechRepublic/1848K9/2009_dom_LP

Table C1: Sequences in avian H7 HA post-1990 Eurasian dataset. Sequences are labelled in format “Serotype”_ “year of sampling”_ “NCBI sequence name”_ “avian host (wild or domestic)”_“viral pathogenicity (highly pathogenic = HP, low pathogenic = LP)”.

Trait transition model	Joint phylogeny and discrete trait sampling, or use of empirical tree distribution
Symmetric CTMC with BSSVS	Empirical tree distribution
Asymmetric CTMC with BSSVS	Empirical tree distribution
Asymmetric CTMC with BSSVS	Joint phylogeny and discrete trait inference
Symmetric CTMC with MJ counting	Empirical tree distribution
Asymmetric CTMC with MJ counting	Empirical tree distribution
Asymmetric CTMC with MJ counting	Joint phylogeny and discrete trait inference

Table C2
Discrete ancestral trait mapping analyses performed on the Eurasian (post 1990) avian H7 HA dataset. Discrete trait evolution across the BEAST phylogenies was modelled using both symmetric and asymmetric continuous time Markov chains (CTMC). The discrete traits considered were viral NA subtype, avian host and viral pathogenicity. In one set of runs, Bayesian stochastic search variable selection (BSSVS) was implemented to identify a parsimonious diffusion model for the dissemination of discrete traits across the phylogeny. In runs where BSSVS was not implemented, ‘Markov jumps’ (MJ) counting was used to record the number of transitions across the tree.

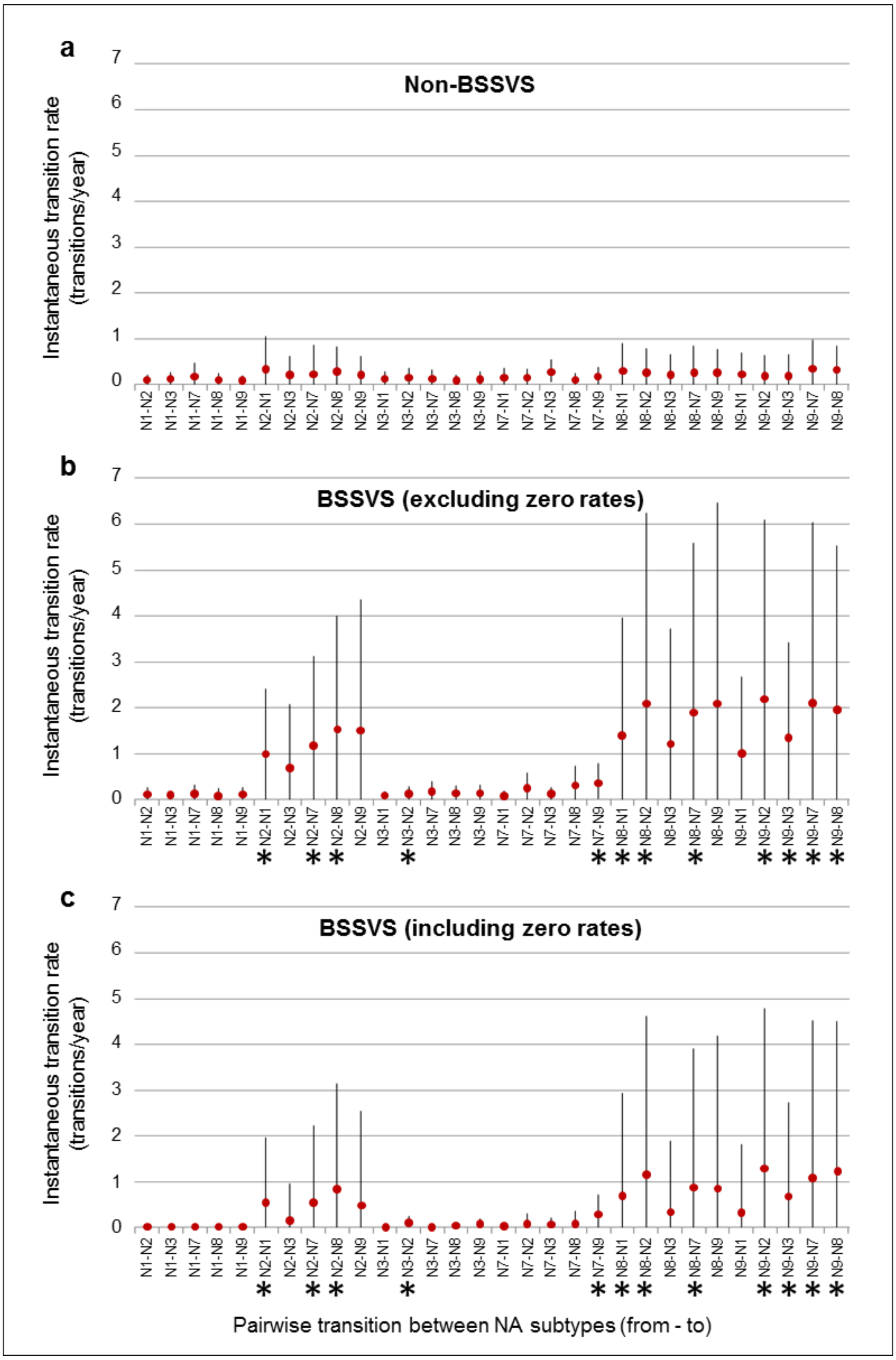


Figure C2: Instantaneous transitions rates between pairs of background NA subtypes. Rates were calculated by multiplying the overall transition clock rate for each MCMC sample by the corresponding relative instantaneous transition rate parameter from the non-BSSVS or BSSVS analyses, (a) excluding 'switched-off' transition rates with an indicator value of zero and (b) including 'switched-off' rates. The much larger means and variances for the BSSVS runs led to higher means and upper 95% HPD limits for the BSSVS runs compared to the non-BSSVS runs.

Run	Number of states	Number of possible rates	Transition clock rate	Clock rate * tree length	Observed number of jumps
Subtype sym	6	15	0.1385 (0.0663, 0.2248)	36.36 (16.81, 58.96)	33.82 (25, 43)
Subtype asym	6	30	0.183 (0.0715, 0.3257)	48.12 (18.70, 86.51)	35.65 (25, 49)
Host sym	2	1	0.0771 (0.0377, 0.1245)	20.25 (9.59, 32.64)	19.11 (14, 27)
Host asym	2	2	0.0763 (0.0358, 0.1217)	20.08 (9.25, 31.92)	18.39 (13, 25)
Path sym	2	1	0.02056 (0.00432, 0.03902)	5.3989 (1.20, 10.30)	4.395889 (4, 6)
Path asym	2	2	0.0183 (0.00296, 0.0392)	4.839145 (0.868, 10.500)	4.283111 (4, 6)

Table C3
Relationship between expected and observed number of state transitions across tree samples. The observed number of discrete trait transitions across the tree from a Markov jumps analysis was compared with the predicted number of transitions obtained by multiplying the overall transition clock rate by the corresponding tree length (also see Appendix C, Figure C3). Means and 95% HPD intervals from across the MCMC are reported here.

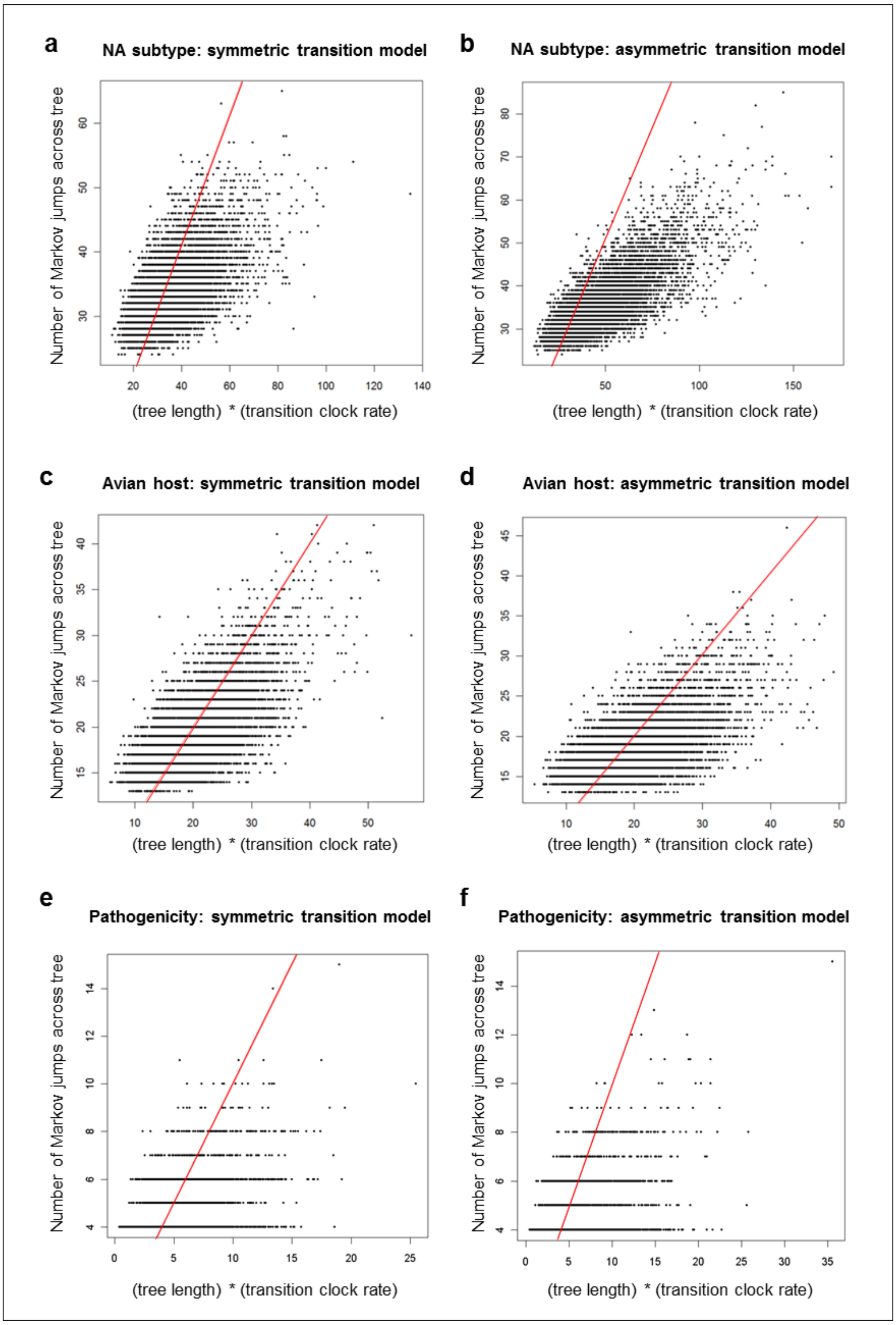


Figure C3

Comparing the expected number of discrete trait transitions with the number of Markov jumps across phylogenies. For each phylogeny sample, an overall transition clock rate was calculated for transition between NA subtypes, avian hosts and viral pathogenicity. The expected number of transitions across the tree was calculated by multiplying the transition clock rate by the total length of the tree. This was plotted against the total number of discrete trait transitions from a Markov jumps analysis, in order to compare the methods. Lines with the equation $y = x$ are also plotted, to assist in assessing the concordance between the clock rates and Markov jump counts.

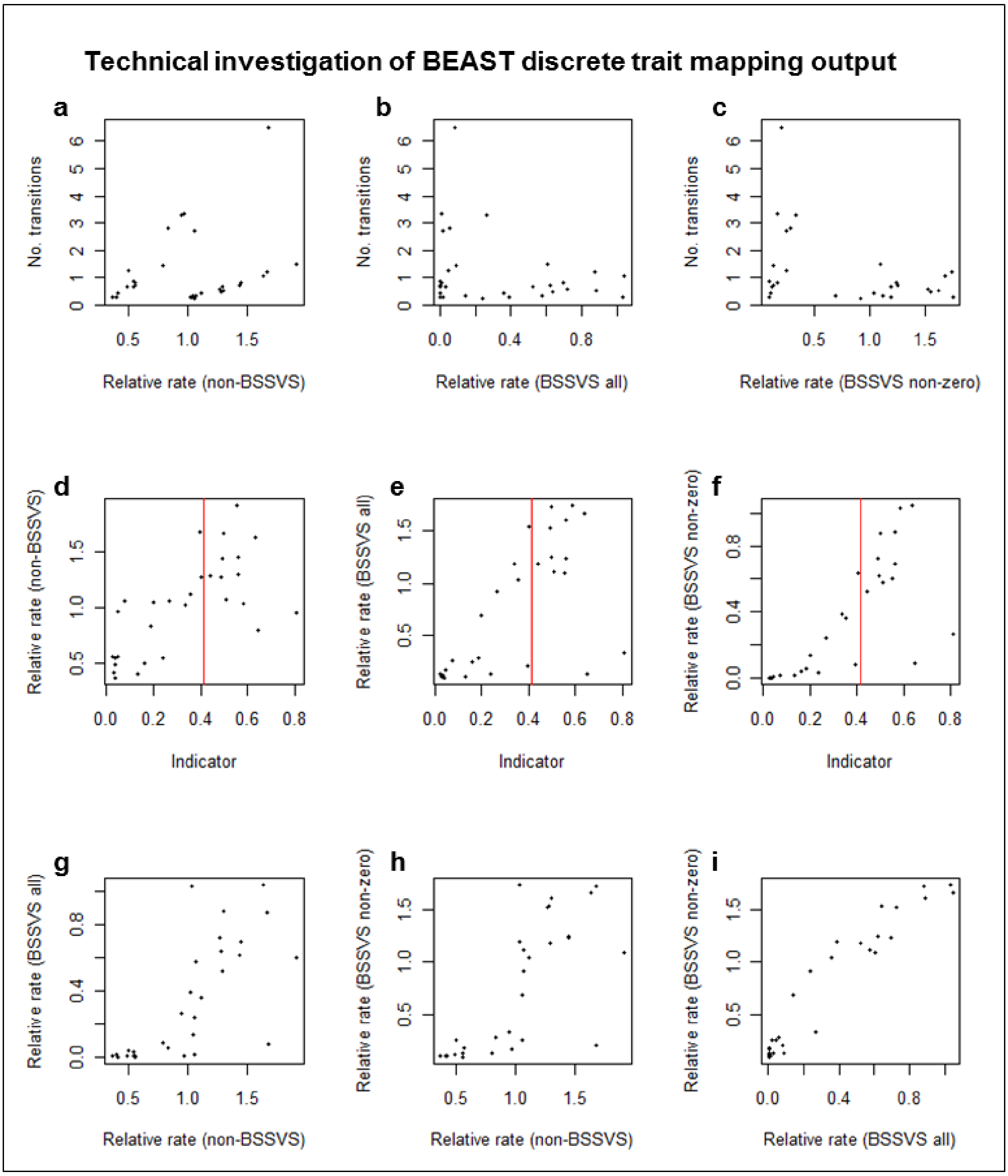


Figure C4

Relationship between outputs of BEAST discrete trait mapping analyses. For the mapping of viral NA subtype onto avian H7 HA phylogeny samples, no obvious relationship was found between the number of pair-wise NA subtype transitions and the relative rate parameters for those transitions (plots a-c; Appendix C, Table C4). A positive correlation was observed between the relative instantaneous rate parameters (from BSSVS and non-BSSVS analyses – see main text for details on calculation) and the indicator value describing the proportion of the time a particular rate was switched on (plots d-f; Appendix C, Table C4). Positive correlations were also found between all different pairs of ways of calculating the relative rate parameters (plots g-i; Appendix C, Table C4). The red vertical line in plots d-f indicates the required cut-off value for a rate to be ‘significantly non-zero’ (Bayes factor > 3).

Output 1	Output 2	Spearman's rho	p -value (uncorrected)	Significance after Bonferroni correction (p <0.0056)
Relative rate (non BSSVS)	Number of transitions	0.2400	0.2006	
Relative rate (BSSVS all)	Number of transitions	-0.0621	0.7439	
Relative rate (BSSVS non-zero)	Number of transitions	-0.0714	0.7068	
Indicator	Relative rate (non BSSVS)	0.6716	$7.32\text{x}10^{-5}$	*
Indicator	Relative rate (BSSVS non-zero)	0.7335	$7.89\text{x}10^{-6}$	*
Indicator	Relative rate (BSSVS all)	0.8692	$5.29\text{x}10^{-7}$	*
Relative rate (non BSSVS)	Relative rate (BSSVS all)	0.7944	$1.35\text{x}10^{-6}$	*
Relative rate (non BSSVS)	Relative rate (BSSVS non-zero)	0.7900	$1.45\text{x}10^{-6}$	*
Relative rate (BSSVS all)	Relative rate (BSSVS non-zero)	0.9511	$< 2.2\text{x}10^{-16}$	*

Table C4

Relationship between BEAST discrete trait mapping outputs. Spearman’s rank correlation was calculated between different parameters from the BEAST discrete trait mapping of viral NA subtypes onto H7 HA phylogeny samples (note that Pearson’s correlation was not appropriate due to the possibility of non-normality in the 30 data-points). For an asymmetric model with 6 states, a total of 30 (=6*5) pairwise transition rates were considered. The mean number of transitions from one state to another, the indicator (proportion of MCMC states at which the particular state was switched on) and the mean instantaneous relative rate parameter describing how often transition from a given state to another particular state occurs with respect to the other pairwise transitions were considered. Relative rate parameters were calculated from a non-BSSVS analysis; in BSSVS analyses the relative rates were multiplied by the indicator values at each sampled MCMC state (‘BSSVS all’), and in some analyses only states with non-zero indicators were included (‘BSSVS non-zero’). A Bonferroni correction was performed to determine the required *p*-value cut-off (0.05/9 = 0.0056) given that 9 comparisons were being performed. All previously significant correlations remained significant under the Bonferroni correction.

10.4 Appendix D

Sequence name	p17	gp41	p17_labels_i	gp41_lables_i	p17_labels_ii	gp41_labels_ii
CD_31265_1984	A	CRF01	A	CRF01	A	CRF01
CD_31299_1984	A	A	A	A	A	A
CD_31331_1984	D	D	D	D	D	D
CD_31335_1984	A	H	A	H	A	H
CD_31496_1984	A	A	A	A	A	A
CD_31518_1984	A	A	A	A	A	A
CD_31533_1984	D	D	D	D	D	D
CD_31598_1984	A	G	A	G	A	G
CD_31602_1984	C	C	C	C	C	C
CD_31605_1984	D	G	D	G	D	G
CD_31653_1984	D	D	D	D	D	D
CD_31688_1984	D	D	D	D	D	D
CD_31748_1984	A	A	A	A	A	A
CD_31807_1984	A	J	A	J	A	J
CD_31821_1984	D	F	D	F	D	F
CD_31857_1984	F	F	F	F	F	F
CD_31873_1984	H	A	H	A	H	A
CD_31886_1984	A	A	A	A	A	A
CD_31899_1984	G	G	G	G	G	G
CD_31917_1984	J	U1	J	G	J	G
CD_31978_1984	A	A	A	A	A	A
CD_32051_1984	A	J	A	J	A	J
CD_32128_1984	D	D	D	D	A	A
CD_32130_1984	D	D	D	D	A	A
CD_32154_1984	A	CRF01	A	CRF01	A	CRF01
CD_32170_1984	A	A	A	A	A	A
CD_32188_1984	D	F1	D	F	D	F
CD_32290_1984	G	G	G	G	G	G
CD_30008_1984	G	G	G	G	G	G
CD_30084_1984	G	G	G	G	G	G
CD_30104_1984	A	A	A	A	A	A
CD_30105_1984	A	A	A	A	A	A
CD_30109_1984	A	A	A	A	A	A
CD_30146_1984	K	K	K	K	K	K
CD_30184_1984	D	D	D	D	D	D
CD_30280_1984	J	J	J	J	J	J
CD_30326_1984	C	C	C	C	C	C
CD_30407_1984	A	A	A	A	A	A
CD_30432_1984	D	D	D	D	D	D
CD_30487_1984	A	CRF01	A	CRF01	A	CRF01
CD_30506_1984	A	G	A	G	A	G

Continued on next page

CD_30509_1984	F	F	F	F	F	F
CD_30582_1984	G	G	G	G	G	G
CD_30750_1984	A	G	A	G	A	G
CD_30793_1984	F	F	F	F	F	F
CD_30871_1984	A	A	A	A	A	A
CD_30873_1984	G	A	G	A	G	A
CD_30884_1984	U2	U2	D	D	U	U
CD_31149_1984	G	G	G	G	G	G
CD_31166_1984	D	D	D	D	D	D
CD_31184_1984	K	K	K	K	K	K
CD_31194_1984	A	A	A	A	A	A
CD_31252_1984	D	D	D	D	D	D

Table D1
Subtype labels assigned to HIV-1 group M sequences. p17 and gp41 sequences were matched by the patient identifier from Kalish *et al*. (2004), and sequence names took the form "two letter country code_ patient identifier_year". The p17 and gp41 sequences were independently assigned subtypes, using maximum likelihood phylogenies which included reference sequences of different subtypes. Subtypes assigned for each region from each individual are reported in the table as "p17" and "gp41". To investigate the possible confounding effect of misclassifying sequences which were difficult to subtype, the inter-subtype recombination rate analysis was repeated twice using alternative labellings ("labels_i" and "labels_ii"). Highlighted rows indicate the sequences which are labelled differently between the original subtype labellings and "labels_i" and/or "labels_ii".

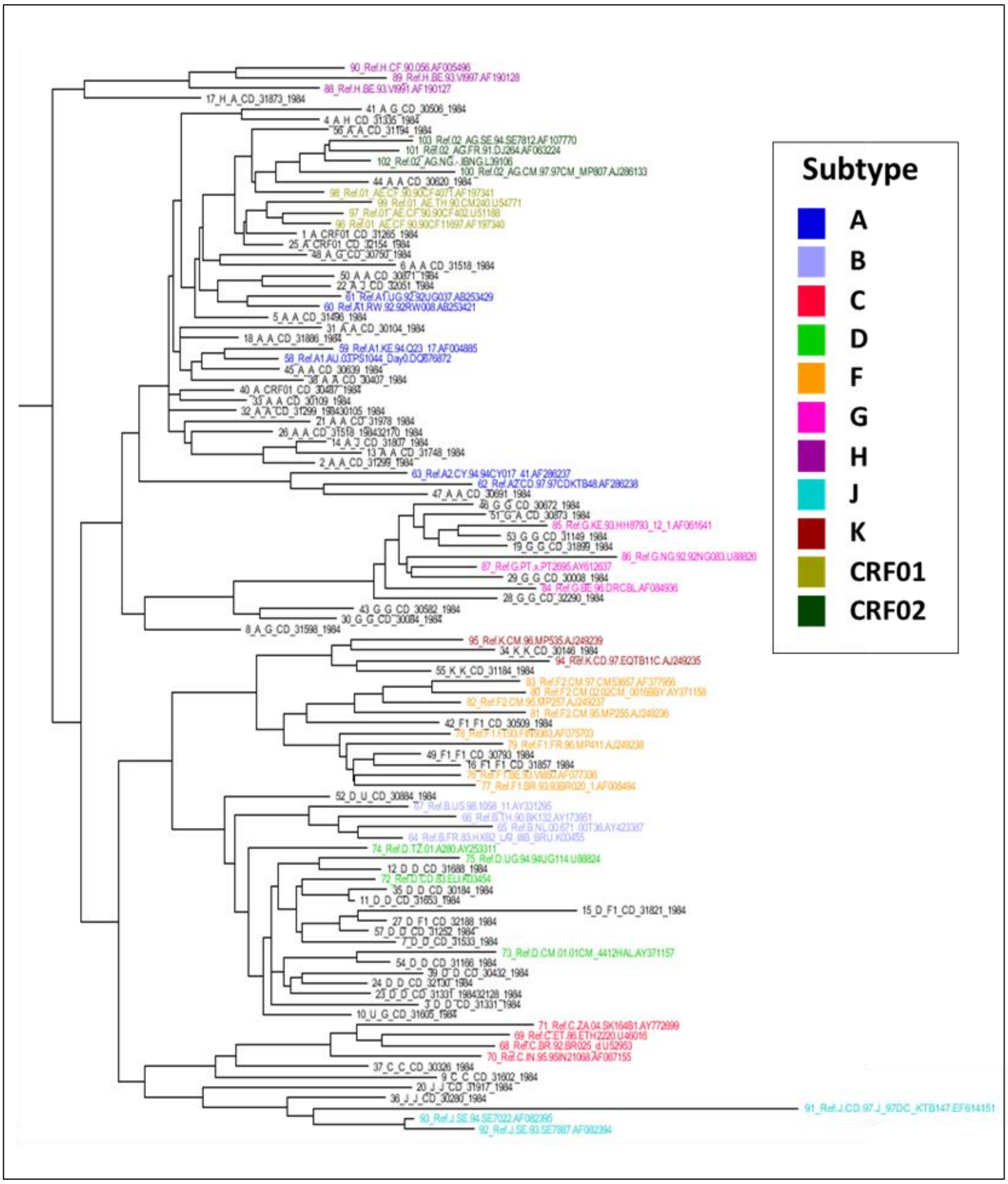

90_Ref.H.CF.90.056.AF005496
89_Ref.H.BE.93.VI997.AF190128
88_Ref.H.BE.93.VI991.AF190127
17_H_A_CD_31873_1984
41_A_G_CD_30506_1984
4_A_H_CD_31335_1984
56_A_A_CD_31194_1984
103_Ref.02_AG.SE.94.SE7812.AF107770
101_Ref.02_AG.FR.91.DJ264.AF063224
102_Ref.02_AG.NG.-.IBNG.L39106
100_Ref.02_AG.CM.97.97CM_MP807.AJ286133
44_A_A_CD_30620_1984
98_Ref.01_AE.CF.90.90CF4071.AF197341
99_Ref.01_AE.TH.90.CM240.U54771
97_Ref.01_AE.CF.90.90CF402.U51188
96_Ref.01_AE.CF.90.90CF11697.AF197340
1_A_CRF01_CD_31265_1984
25_A_CRF01_CD_32154_1984
48_A_G_CD_30750_1984
6_A_A_CD_31518_1984
50_A_A_CD_30871_1984
22_A_J_CD_32051_1984
61_Ref.A1.UG.92.92UG037.AB253429
60_Ref.A1.RW.92.92RW008.AB253421
5_A_A_CD_31496_1984
31_A_A_CD_30104_1984
18_A_A_CD_31886_1984
59_Ref.A1.KE.94.Q23_17.AF004885
58_Ref.A1.AU.03.PS1044_Day0.DQ676872
45_A_A_CD_30639_1984
38_A_A_CD_30407_1984
40_A_CRF01_CD_30487_1984
33_A_A_CD_30109_1984
32_A_A_CD_31299_198430105_1984
21_A_A_CD_31978_1984
26_A_A_CD_31518_198432170_1984
14_A_J_CD_31807_1984
13_A_A_CD_31748_1984
2_A_A_CD_31299_1984
63_Ref.A2.CY.94.94CY017_41.AF286237
62_Ref.A2.CD.97.97CDKTB48.AF286238
47_A_A_CD_30691_1984
46_G_G_CD_30672_1984
51_G_A_CD_30873_1984
85_Ref.G.KE.93.HH8793_12_1.AF061641
53_G_G_CD_31149_1984
19_G_G_CD_31899_1984
86_Ref.G.NG.92.92NG083.U88826
87_Ref.G.PT.x.PT2695.AY612637
29_G_G_CD_30008_1984
84_Ref.G.BE.96.DRCBL.AF084936
28_G_G_CD_32290_1984
43_G_G_CD_30582_1984
30_G_G_CD_30084_1984
8_A_G_CD_31598_1984
95_Ref.K.CM.96.MP535.AJ249239
34_K_K_CD_30146_1984
94_Ref.K.CD.97.EQTB11C.AJ249235
55_K_K_CD_31184_1984
83_Ref.F2.CM.97.CM53657.AF377956
80_Ref.F2.CM.02.02CM_0016BBY.AY371158
82_Ref.F2.CM.95.MP257.AJ249237
81_Ref.F2.CM.95.MP255.AJ249236
42_F1_F1_CD_30509_1984
78_Ref.F1.FI.93.FIN9363.AF075703
79_Ref.F1.FR.96.MP411.AJ249238
49_F1_F1_CD_30793_1984
16_F1_F1_CD_31857_1984
76_Ref.F1.BE.93.VI850.AF077336
77_Ref.F1.BR.93.93BR020_1.AF005494
52_D_U_CD_30884_1984
67_Ref.B.US.98.1058_11.AY331295
66_Ref.B.TH.90.BK132.AY173951
65_Ref.B.NL.00.671_00T36.AY423387
64_Ref.B.FR.83.HXB2_LAI_IIIB_BRU.K03455
74_Ref.D.TZ.01.A280.AY253311
75_Ref.D.UG.94.94UG114.U88824
12_D_D_CD_31688_1984
72_Ref.D.CD.83.ELI.K03454
35_D_D_CD_30184_1984
11_D_D_CD_31653_1984
15_D_F1_CD_31821_1984
27_D_F1_CD_32188_1984
57_D_D_CD_31252_1984
7_D_D_CD_31533_1984
73_Ref.D.CM.01.01CM_4412HAL.AY371157
54_D_D_CD_31166_1984
39_D_D_CD_30432_1984
24_D_D_CD_32130_1984
23_D_D_CD_31331_198432128_1984
3_D_D_CD_31331_1984
10_U_G_CD_31605_1984
71_Ref.C.ZA.04.SK164B1.AY772699
69_Ref.C.ET.86.ETH2220.U46016
68_Ref.C.BR.92.BR025_d.U52953
70_Ref.C.IN.95.95IN21068.AF067155
37_C_C_CD_30326_1984
9_C_C_CD_31602_1984
20_J_J_CD_31917_1984
36_J_J_CD_30280_1984
91_Ref.J.CD.97.J_97DC_KTB147.EF614151
93_Ref.J.SE.94.SE7022.AF082395
92_Ref.J.SE.93.SE7887.AF082394
Subtype
A
B
C
D
F
G
H
J
K
CRF01
CRF02

Figure D1
Maximum likelihood tree for HIV-1 group M p17 region. A maximum likelihood phylogeny was constructed using PHYML, under a general time-reversible model of nucleotide substitution with gamma-distributed rate heterogeneity across sites (alpha parameter estimated from the data) and 6 rate categories. The data comprised p17 sequences reported by Kalish *et al*. (2004) for which a corresponding gp41 sequence from that individual was available, as well as reference sequences for all HIV-1 group M subtypes, and CRFs 01 and 02, from the Los Alamos National Laboratory database. 1000 bootstrap replicates were performed (bootstrap values not shown). The tree was rooted to a p17 sequence from HIV-1 group N.

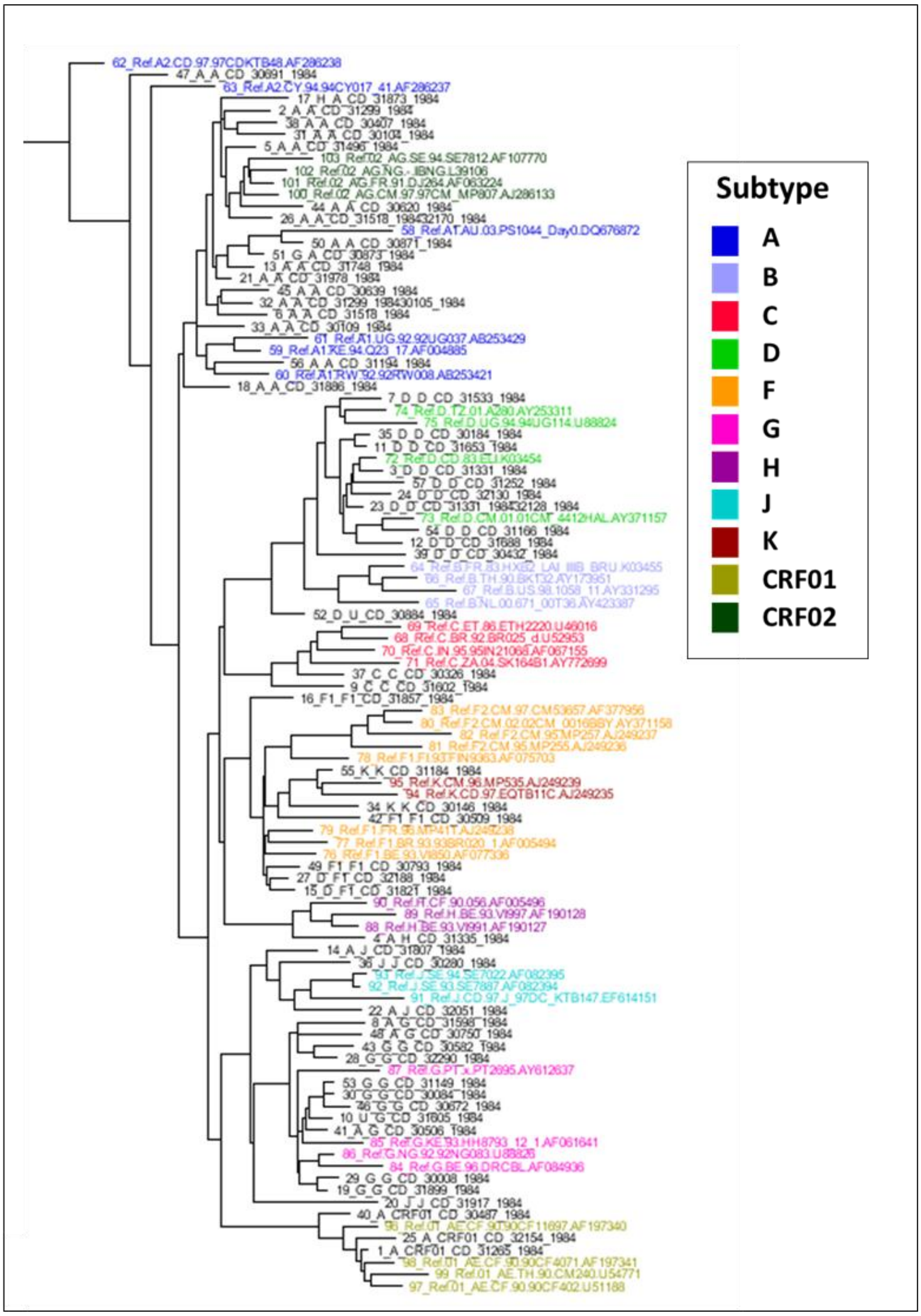

62_Ref.A2.CD.97.97CDKTB48.AF286238
47_A_A_CD_30691_1984
63_Ref.A2.CY.94.94CY017_41.AF286237
17_H_A_CD_31873_1984
2_A_A_CD_31299_1984
38_A_A_CD_30407_1984
31_A_A_CD_30104_1984
5_A_A_CD_31496_1984
103_Ref.02_AG.SE.94.SE7812.AF107770
102_Ref.02_AG.NG.-.IBNG.L39106
101_Ref.02_AG.FR.91.DJ264.AF063224
100_Ref.02_AG.CM.97.97CM_MP807.AJ286133
44_A_A_CD_30620_1984
26_A_A_CD_31518_198432170_1984
58_Ref.AT.AU.03.PS1044_Day0.DQ676872
50_A_A_CD_30871_1984
51_G_A_CD_30873_1984
13_A_A_CD_31748_1984
21_A_A_CD_31978_1984
45_A_A_CD_30639_1984
32_A_A_CD_31299_198430105_1984
6_A_A_CD_31518_1984
33_A_A_CD_30109_1984
61_Ref.A1.UG.92.92UG037.AB253429
59_Ref.A1.KE.94.Q23_17.AF004885
56_A_A_CD_31194_1984
60_Ref.A1.RW.92.92RW008.AB253421
18_A_A_CD_31886_1984
7_D_D_CD_31533_1984
74_Ref.D.TZ.01.A280.AY253311
75_Ref.D.UG.94.94UG114.U88824
35_D_D_CD_30184_1984
11_D_D_CD_31653_1984
72_Ref.D.CD.83.ELI.K03454
3_D_D_CD_31331_1984
57_D_D_CD_31252_1984
24_D_D_CD_32130_1984
23_D_D_CD_31331_198432128_1984
73_Ref.D.CM.01.01CM_4412HAL.AY371157
54_D_D_CD_31166_1984
12_D_D_CD_31688_1984
39_D_D_CD_30432_1984
64_Ref.B.FR.83.HXB2_LAI_IIIB_BRU.K03455
66_Ref.B.TH.90.BK132.AY173951
67_Ref.B.US.98.1058_11.AY331295
65_Ref.B.NL.00.671_00T36.AY423387
52_D_U_CD_30884_1984
69_Ref.C.ET.86.ETH2220.U46016
68_Ref.C.BR.92.BR025_d.U52953
70_Ref.C.IN.95.95IN21068.AF067155
71_Ref.C.ZA.04.SK164B1.AY772699
37_C_C_CD_30326_1984
9_C_C_CD_31602_1984
16_F1_F1_CD_31857_1984
83_Ref.F2.CM.97.CM53657.AF377956
80_Ref.F2.CM.02.02CM_0016BBY.AY371158
82_Ref.F2.CM.95.MP257.AJ249237
81_Ref.F2.CM.95.MP255.AJ249236
78_Ref.F1.FI.93.FIN9363.AF075703
55_K_K_CD_31184_1984
95_Ref.K.CM.96.MP535.AJ249239
94_Ref.K.CD.97.EQTB11C.AJ249235
34_K_K_CD_30146_1984
42_F1_F1_CD_30509_1984
79_Ref.F1.FR.96.MP411.AJ249238
77_Ref.F1.BR.93.93BR020_1.AF005494
76_Ref.F1.BE.93.VI850.AF077336
49_F1_F1_CD_30793_1984
27_D_F1_CD_32188_1984
15_D_F1_CD_31821_1984
90_Ref.H.CF.90.056.AF005496
89_Ref.H.BE.93.VI997.AF190128
88_Ref.H.BE.93.VI991.AF190127
4_A_H_CD_31335_1984
14_A_J_CD_31807_1984
36_J_J_CD_30280_1984
93_Ref.J.SE.94.SE7022.AF082395
92_Ref.J.SE.93.SE7887.AF082394
91_Ref.J.CD.97.J_97DC_KTB147.EF614151
22_A_J_CD_32051_1984
8_A_G_CD_31598_1984
48_A_G_CD_30750_1984
43_G_G_CD_30582_1984
28_G_G_CD_32290_1984
87_Ref.G.PT.x.PT2695.AY612637
53_G_G_CD_31149_1984
30_G_G_CD_30084_1984
46_G_G_CD_30672_1984
10_U_G_CD_31605_1984
41_A_G_CD_30506_1984
85_Ref.G.KE.93.HH8793_12_1.AF061641
86_Ref.G.NG.92.92NG083.U88826
84_Ref.G.BE.96.DRCBL.AF084936
29_G_G_CD_30008_1984
19_G_G_CD_31899_1984
20_J_J_CD_31917_1984
40_A_CRF01_CD_30487_1984
96_Ref.01_AE.CF.90.90CF11697.AF197340
25_A_CRF01_CD_32154_1984
1_A_CRF01_CD_31265_1984
98_Ref.01_AE.CF.90.90CF4071.AF197341
99_Ref.01_AE.TH.90.CM240.U54771
97_Ref.01_AE.CF.90.90CF402.U51188
Subtype
A
B
C
D
F
G
H
J
K
CRF01
CRF02

Figure D2

Maximum likelihood tree for HIV-1 group M gp41 region. A maximum likelihood phylogeny was constructed using PHYML, under a general time-reversible model of nucleotide substitution with gamma-distributed rate heterogeneity across sites (alpha parameter estimated from the data) and 6 rate categories. The data comprised gp41 sequences reported by Kalish *et al*. (2004) for which a corresponding p17 sequence from that individual was available, as well as reference sequences for all HIV-1 group M subtypes, and CRFs 01 and 02, from the Los Alamos National Laboratory database. 1000 bootstrap replicates were performed (bootstrap values not shown). The tree was rooted to a gp41 sequence from HIV-1 group N.

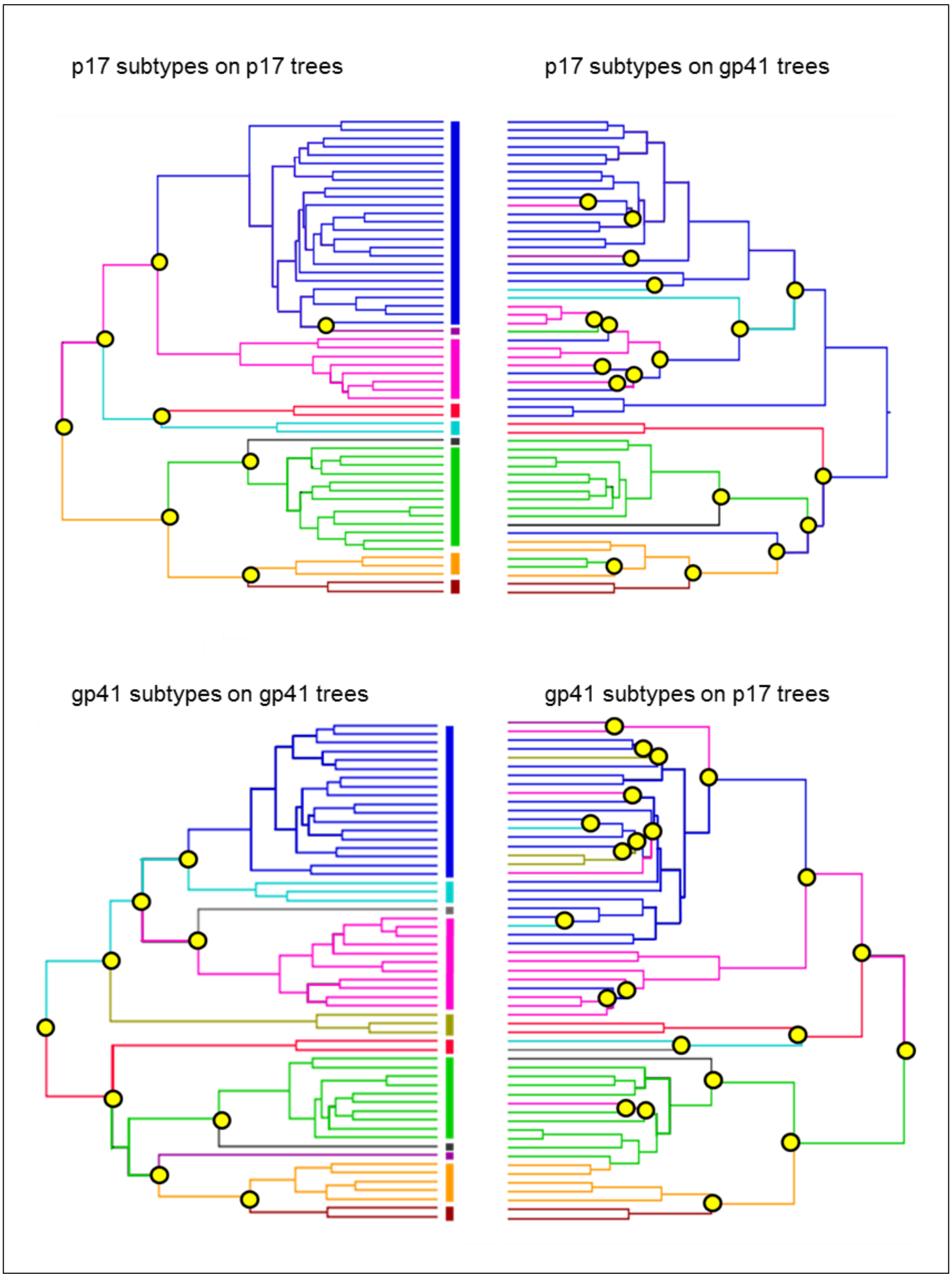

p17 subtypes on p17 trees
p17 subtypes on gp41 trees
gp41 subtypes on gp41 trees
gp41 subtypes on p17 trees

Figure D3
Example of ancestral subtype transitions on the p17 and gp41 phylogeny samples. On each of the example phylogeny samples (gp41 or p17), the branches are coloured according to the inferred p17 or gp41 subtype. Transitions between inferred ancestral subtypes across the tree (observed as colour changes at nodes) are represented by yellow circles. It may be observed that, in this example, a larger number of ancestral transitions were required to map subtypes onto the phylogeny for the correct gene (p17 subtypes onto the p17 tree, or gp41 subtypes onto the gp41 tree) compared to onto the phylogeny for the region at the opposite end of the genome (p17 subtypes onto the gp41 tree, or gp41 subtypes onto the p17 tree).

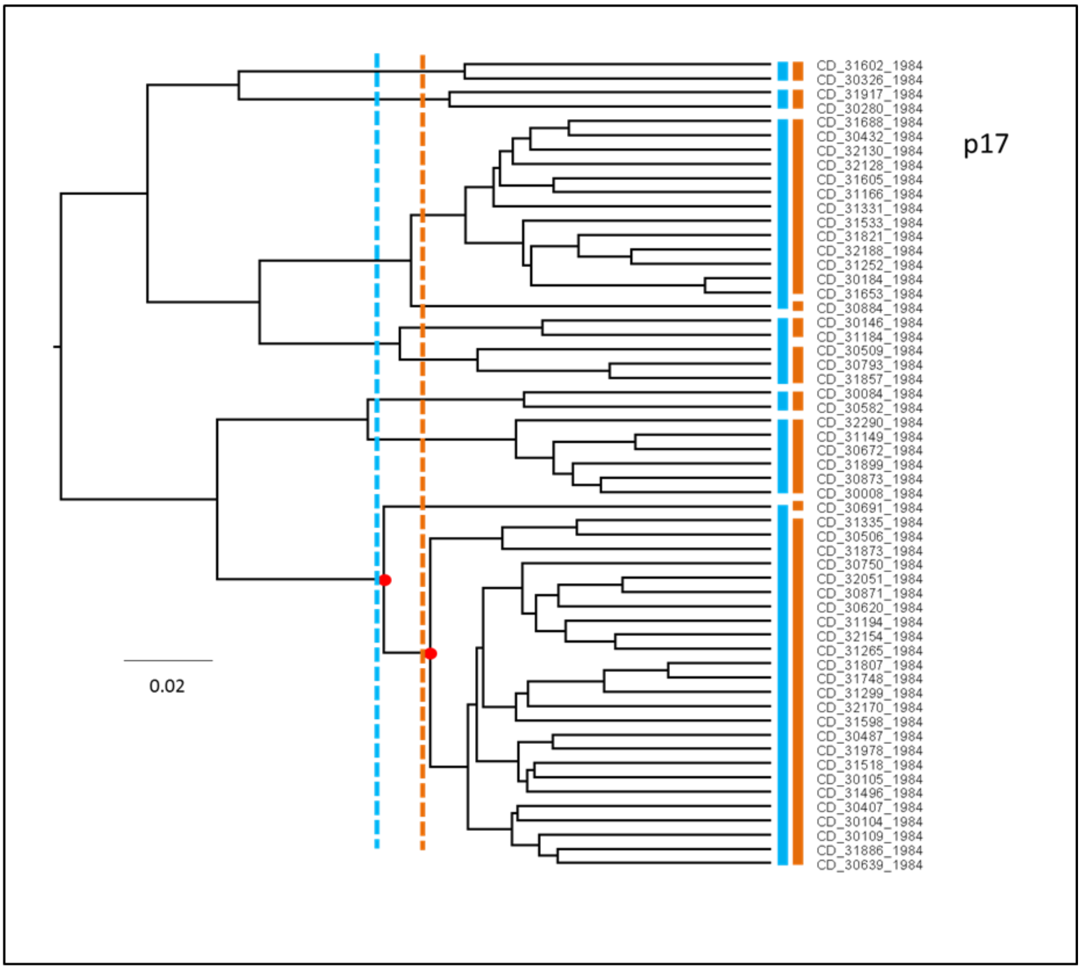


Figure D4
p17 clade labels. To mitigate any error introduced by attempting to classify sequences into subtypes according to the way in which they clustered with the Los Alamos reference sequences, analyses were repeated using clades defined at a specific point along the BEAST maximum clade consensus tree for p17. Two different cut-off points (v1 and v2 respectively) near to the root of the subtype A clade were chosen, which led to the definition of 7 and 10 clades respectively, represented by the blue and orange blocks at the tips of the tree. The scale bar is in units of substitutions per site.

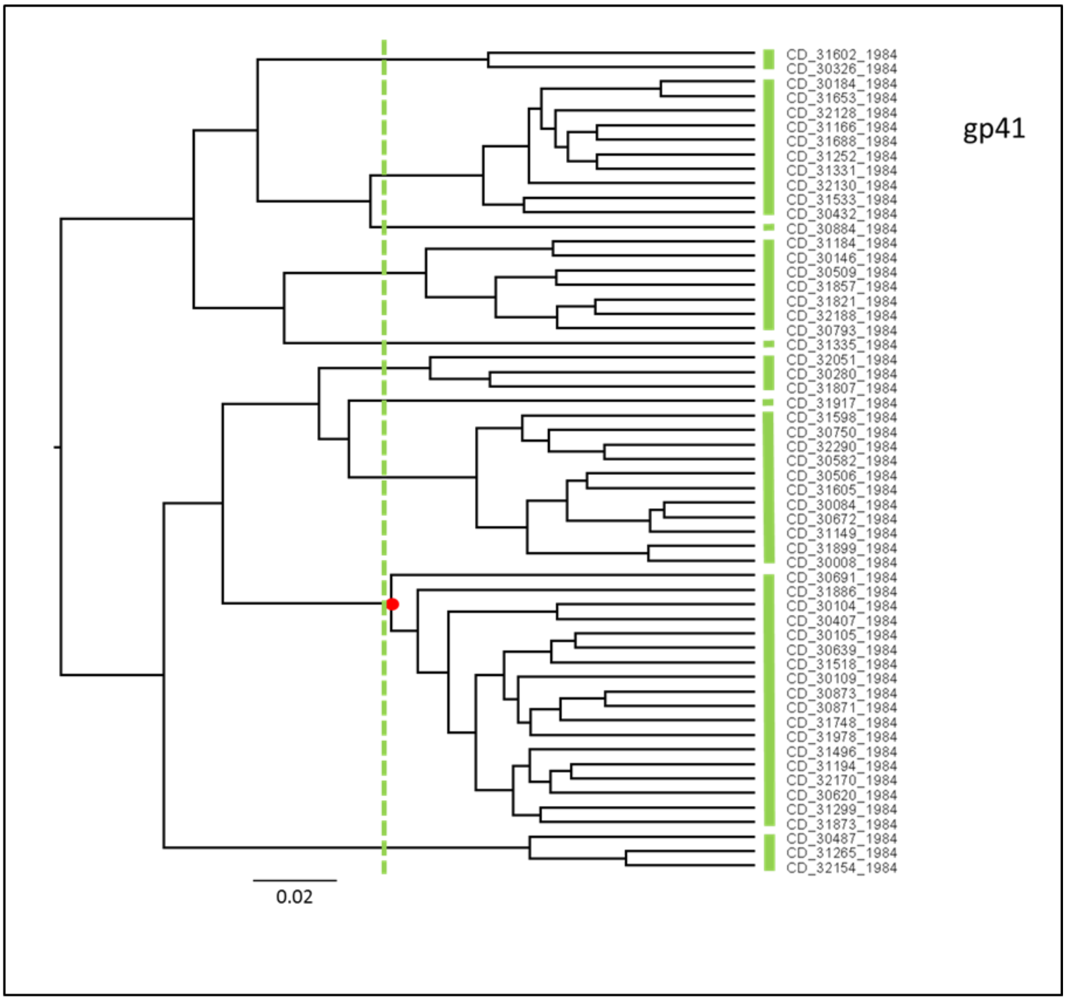


Figure D5
gp41 clade labels. To mitigate any error introduced by attempting to classify sequences into subtypes according to the way in which they clustered with the Los Alamos reference sequences, analyses were repeated using clades defined at a specific point along the BEAST maximum clade consensus tree for gp41. The root of the subtype A clade (represented by a red dot) was chosen as the cut-off. This led to a total of 10 clades being defined (some of which only contained one lineage). The cut-off point is marked on the phylogeny, and the green blocks at the tip of the tree illustrate clades defined in this manner. The scale bar is in units of substitutions per site.

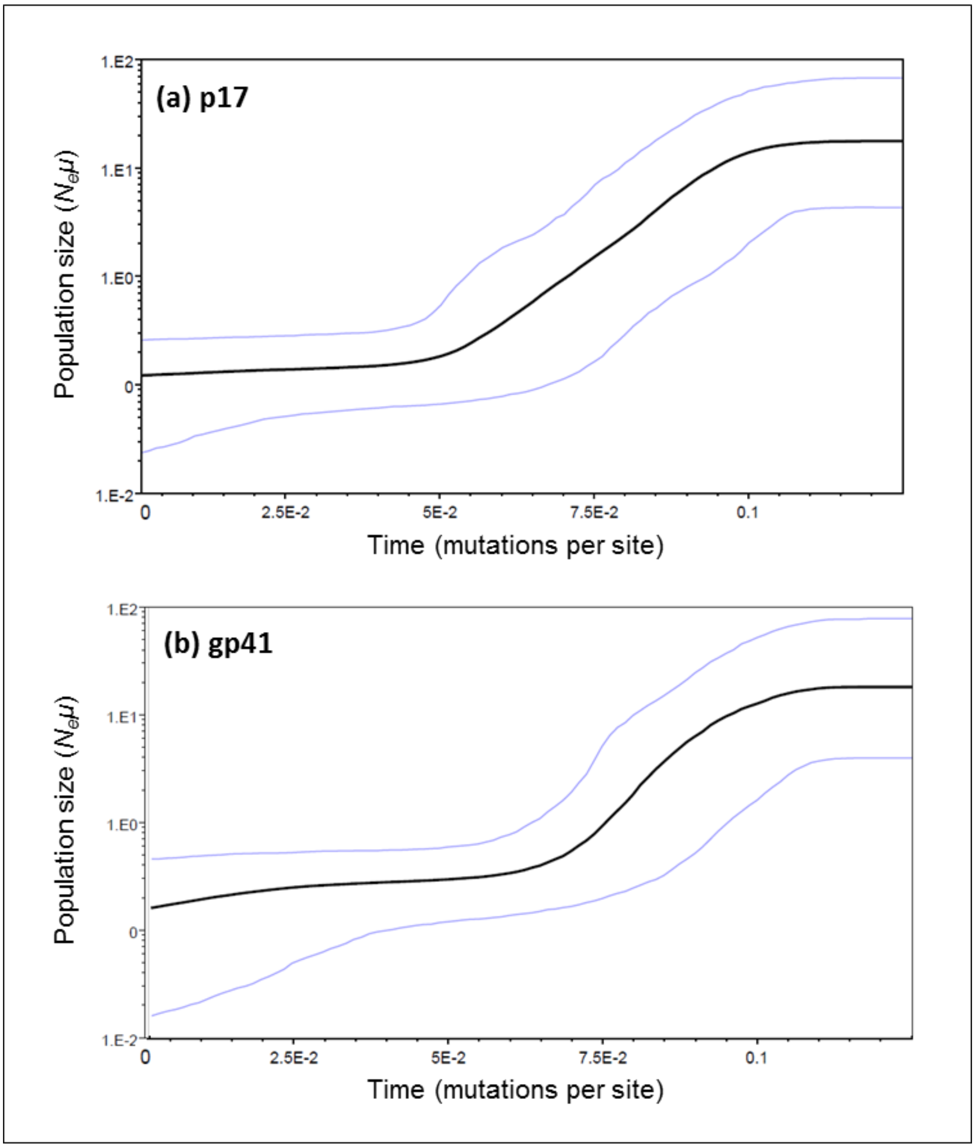


Figure D6
Bayesian skyline plots for p17 and gp41. Coalescent-based analysis was performed in BEAST under a flexible Bayesian skyline demographic model. The plots show the change in viral genetic diversity between the root and tips of the tree, which can be scaled to estimate the effective size of the viral population (and, under certain conditions, the infected population – see Chapter 7). Black lines represent the mean and blue lines represent the 95% HPD limits. The shape of the curve is suggestive of exponential growth, as may be expected across the timescale in which the HIV-1 group M epidemic came to prominence. The apparent rapid increase in viral diversity could also represent underlying changes in the host population, for example demographic expansion, which is consistent with census data for Kinshasa (see Appendix E, Figure E4).

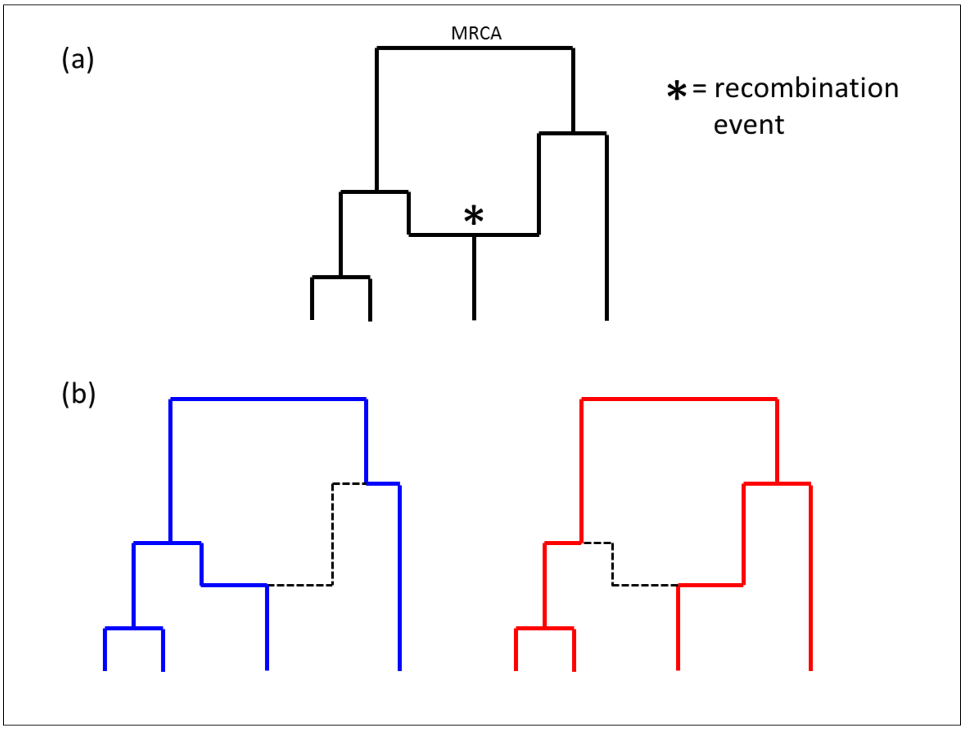


Figure D7
Example ancestral recombination graph (ARG). (a) The ARG shows the history of the sample, moving from the tips of the graph towards the MRCA. In this example, a coalescent event occurs first, then a recombination event, followed by a further three coalescent events until the MRCA has been reached for all sites. (b) The effect of recombination on creating different genealogies for sites on either side of the recombination breakpoint may be observed. The blue lines represent the genealogy for sites on one side of the recombination breakpoint, and the red lines represent the genealogy on the other side. This diagram was based upon Figure 1 of McVean (2001)[5]. Note that the shape of the ARG will depend on the nature and timing of the recombination event in the evolutionary history, as well as the rate of nucleotide substitution, and that not all recombination events which take place will be detectable on an ARG.

[5] McVean, G. A. T. (2001). What do patterns of genetic variability reveal about mitochondrial recombination? *Heredity* **87**: 613-620.

10.5 Appendix E

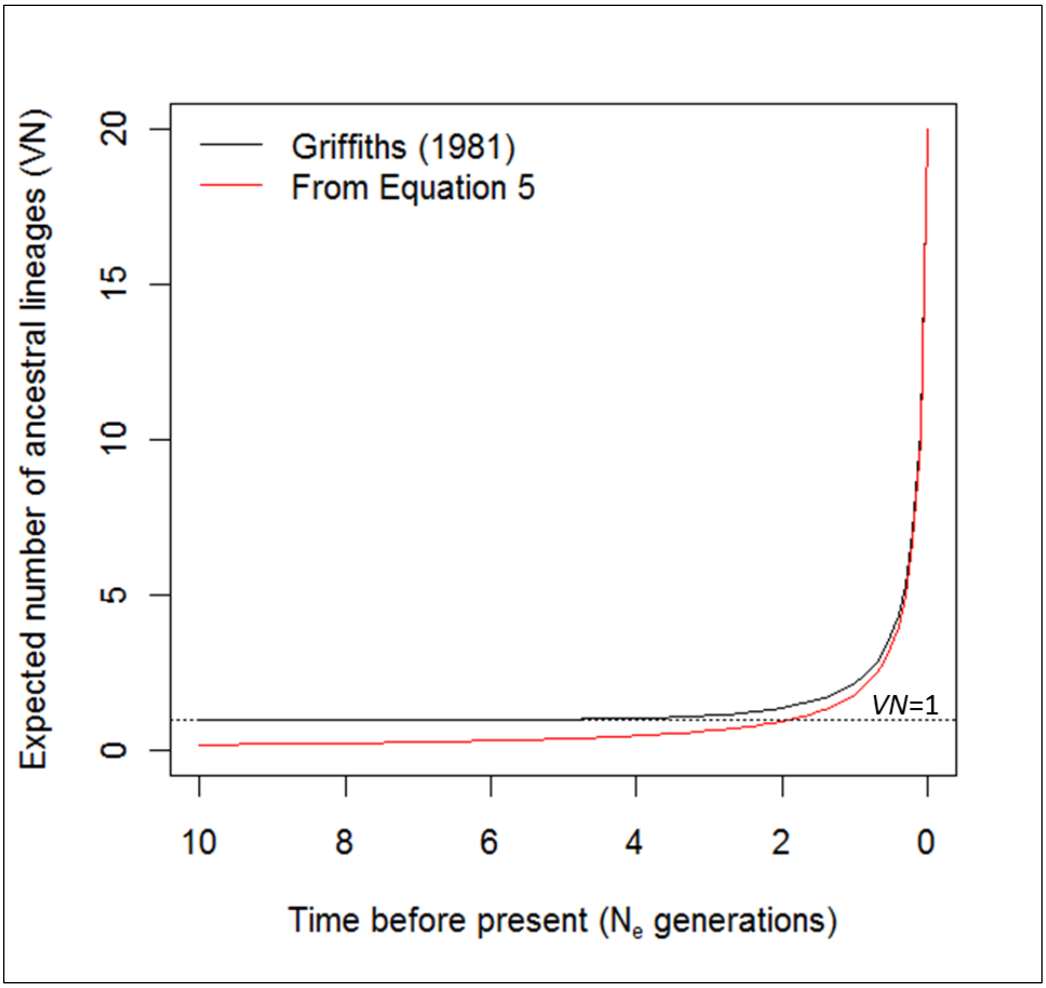


Figure E1
Comparison of the expected number of ancestral lineages (*VN*) under Equation 5 of Chapter 7 and Griffiths (1981). A model of constant population size (equivalent to an endemic model, i.e. $f_{SI}/I^2(t)=N_e/2$) was employed for comparison of the expected number of lineages under Equation 5 (red line), which is an approximation presented by Volz *et al*. (2009), and the exact calculation of the expectation as outlined by Griffiths (1981) and Tavaré (1994) (black line; see equation 6.7 of Tavaré 1984). A sample size of n=20 (number of tips in the tree) was chosen. Time moves from the present (*t*=0) to the past, towards the root of the tree, and is scaled in units of N_e generations. Equation 5 appears to be a reasonable approximation under the assumption of Volz *et al*. (2009) that $V >> 1/N$.

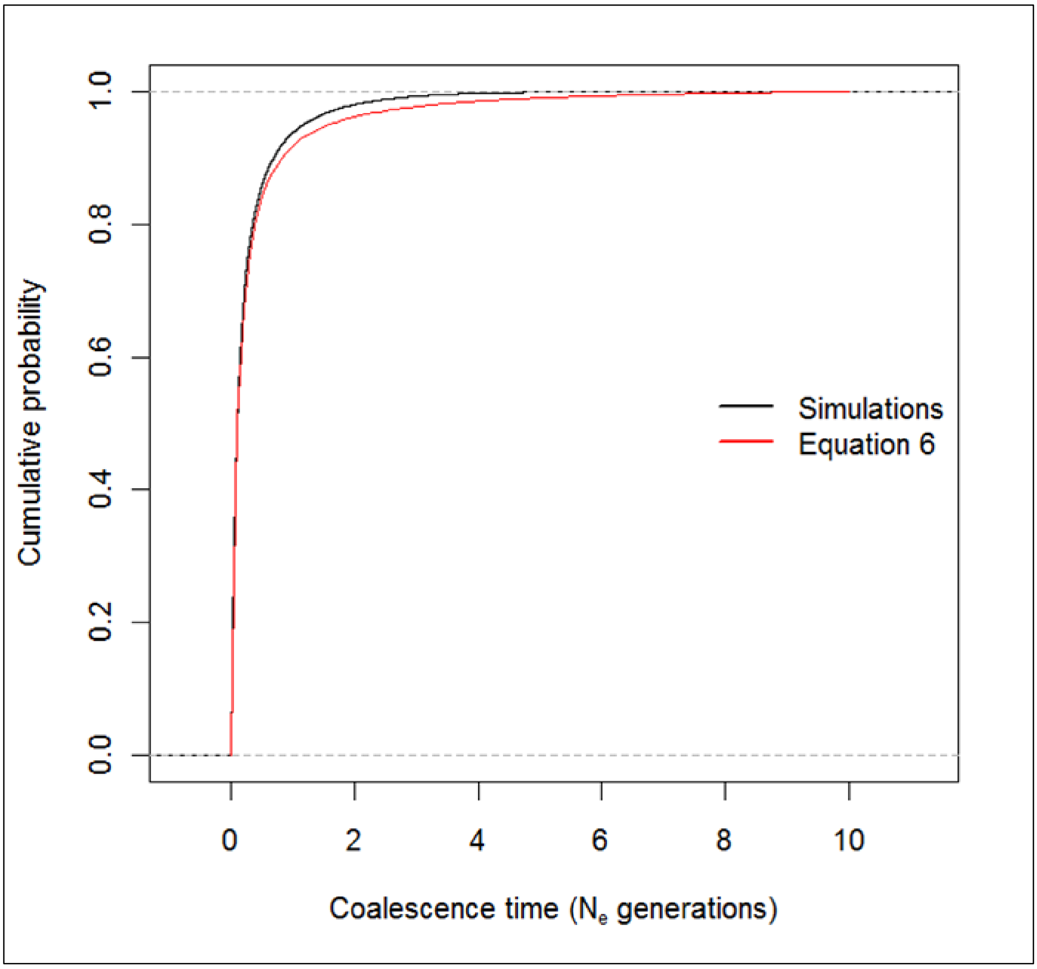


Figure E2
Comparison between Equation 6 and a simulated empirical cumulative distribution function for coalescence times. A model of constant population size was assumed as in Figure E1. For a given number of sampled sequences (here, *n*=20), the times of all *n*-1 coalescence events were simulated randomly from an exponential distribution with rate *kChoose2* when there were *k* lineages (see Section 7.4.2 for a discussion of coalescence rate). The simulation was performed 10,000 times and an empirical cumulative distribution function for the coalescence times was generated and plotted (black line). The empirical CDF could then be compared to the coalescence times under Equation 6 (red line), which appears to be a good approximation to the empirical CDF.

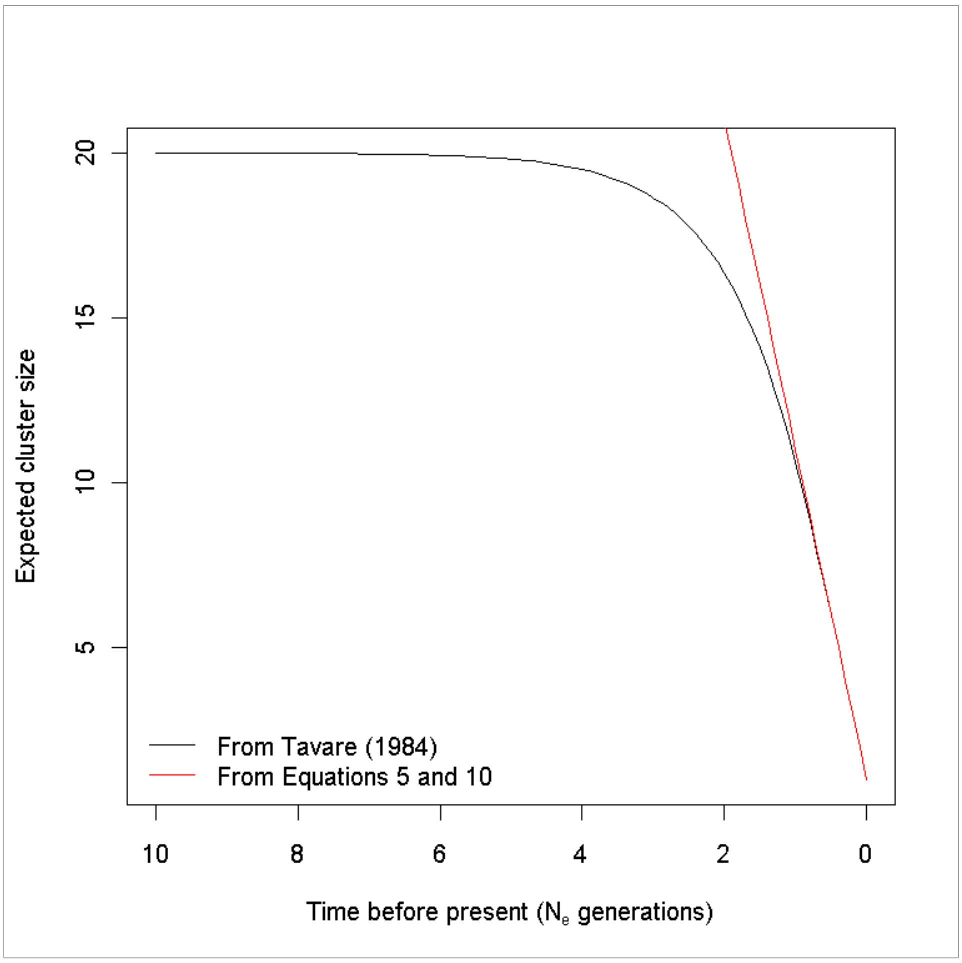


Figure E3
Comparing the approximation for the expected cluster size, x_1, (Chapter 7, Equation 10) with exact results from Tavaré (1984). The expected number of progeny lineages, x_1, at the tips of the tree (time *T*) of a lineage at an earlier time, *t*, was approximated as described in Equation 10 of Chapter 7, under a model of constant population size. This approximation assumed that the number of lineages was much greater than 1 (i.e. that $V >> 1/N$). The approximation to x_1 was plotted against an exact result (using Equations 6.1 and 6.2 of Tavaré (1984), which give the probability that there are *k* lineages at time *t*, given that there are a total of *n* tips in the tree). The approximation appears to be good towards the tips of the tree. However, it does not behave as it should towards the root of the tree (i.e. when the Volz *et al*. (2009) assumption that $V >> 1/N$ no longer holds). Specifically, x_1 should tend to *n* (the number of tips) towards the root of the tree, but the approximation to x_1 carries on increasing beyond *n*.

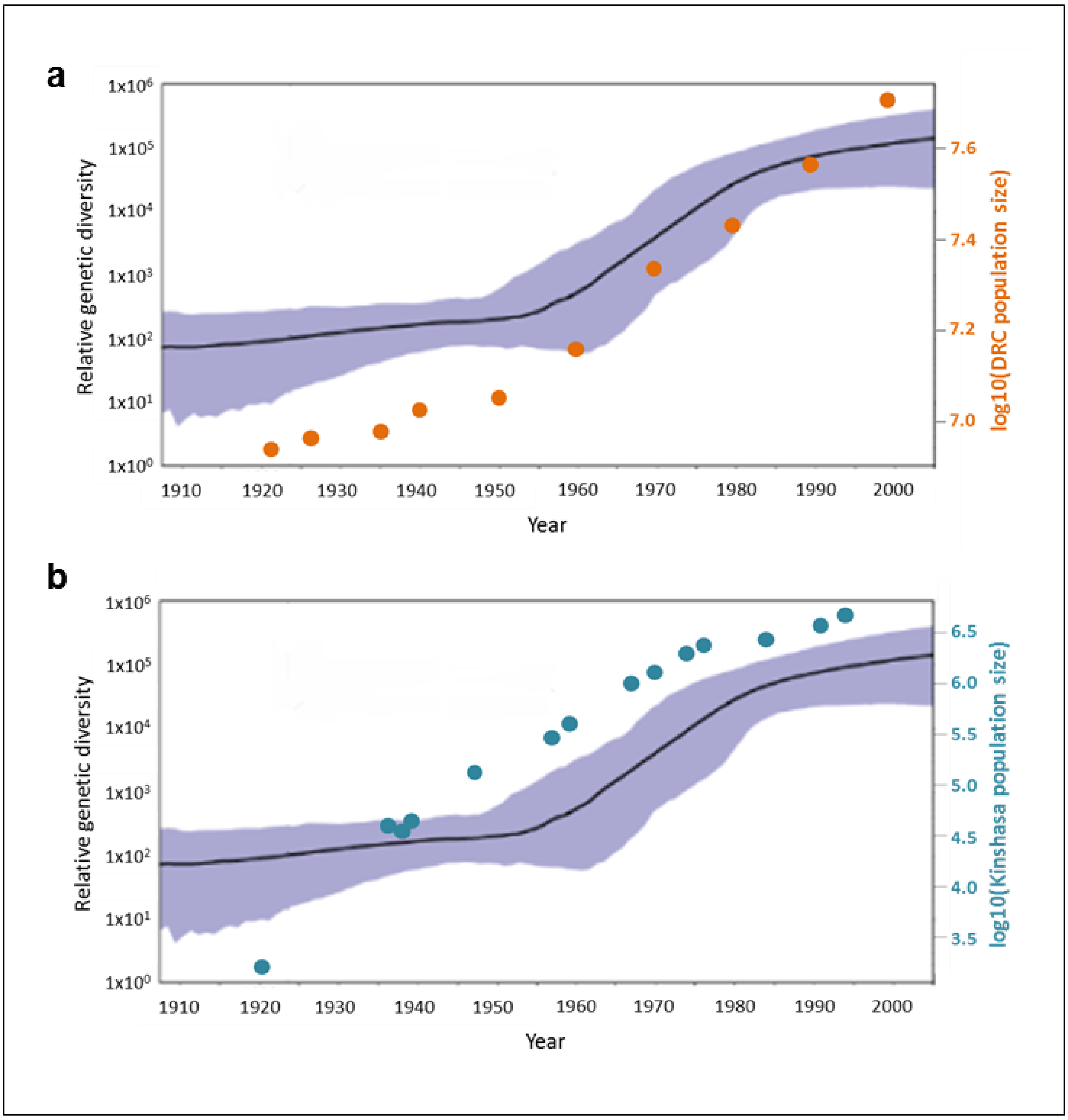
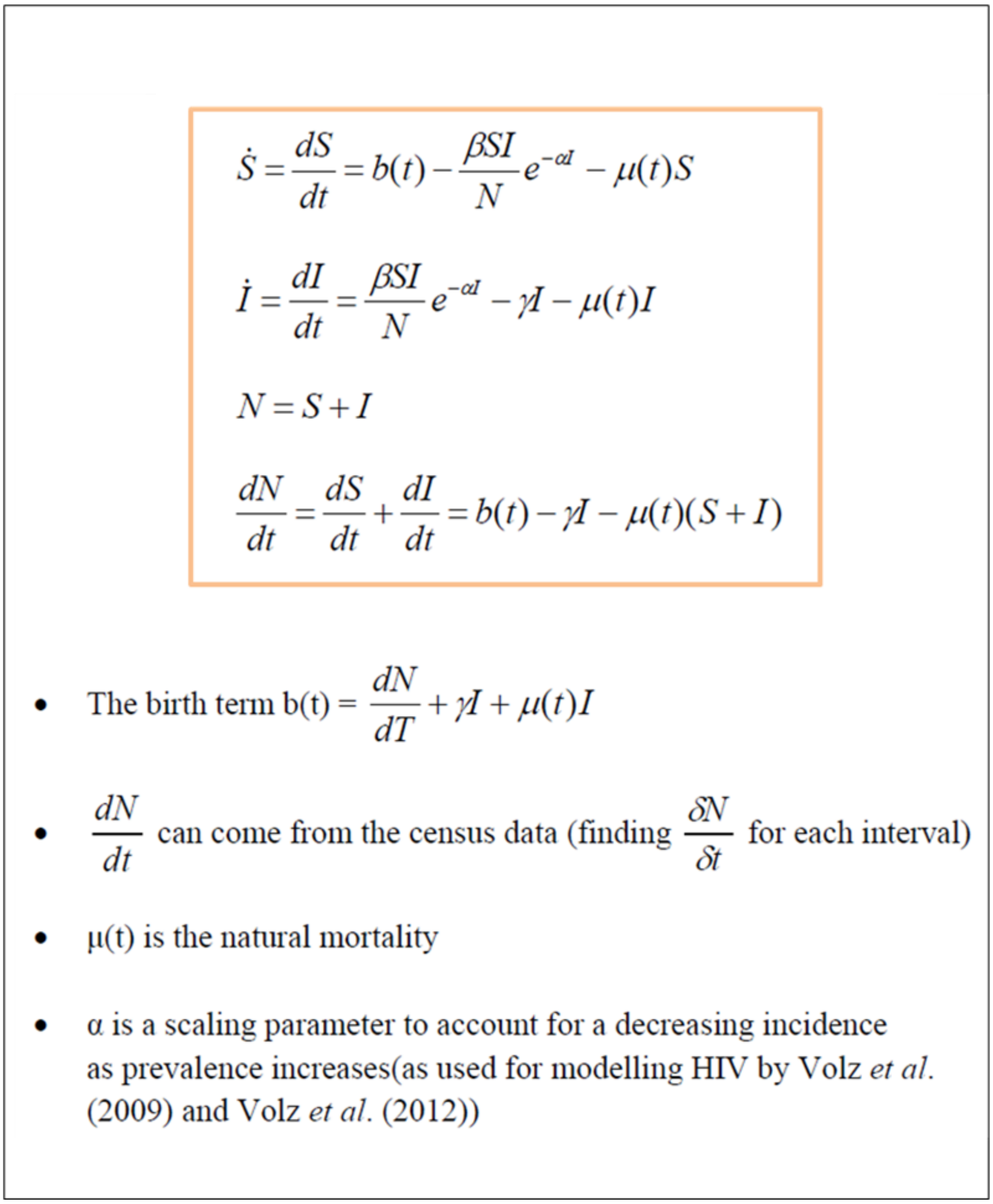


Figure E4

Increase in DRC population size and HIV-1 group M relative genetic diversity over time. Census population size data (coloured circles) were plotted on a logarithmic scale for (a) the Democratic Republic of the Congo (DRC) and (b) Kinshasa, the capital of the DRC. Data were obtained from http://www.populstat.info. The population size data has been superimposed onto Bayesian skyline plots obtained by Worobey *et al*. (2008) for HIV-1 group M sequence data from the DRC, which show a similar pattern of increase in relative genetic diversity to that observed for HIV-1 group M sequences from Kinshasa in Chapter 5 (Appendix D, Figure D6). A similar pattern of increase over time was observed between the relative genetic diversity and both the Kinshasa and DRC population sizes.



$$\dot{S} = \frac{dS}{dt} = b(t) - \frac{\beta SI}{N} e^{-\alpha I} - \mu(t)S$$

$$\dot{I} = \frac{dI}{dt} = \frac{\beta SI}{N} e^{-\alpha I} - \gamma I - \mu(t)I$$

$$N = S + I$$

$$\frac{dN}{dt} = \frac{dS}{dt} + \frac{dI}{dt} = b(t) - \gamma I - \mu(t)(S + I)$$

- The birth term b(t) = $\frac{dN}{dT} + \gamma I + \mu(t)I$
- $\frac{dN}{dt}$ can come from the census data (finding $\frac{\delta N}{\delta t}$ for each interval)
- μ(t) is the natural mortality
- α is a scaling parameter to account for a decreasing incidence as prevalence increases(as used for modelling HIV by Volz *et al*. (2009) and Volz *et al*. (2012))

Figure E5
Compartment model for the HIV-1 group M epidemic in Kinshasa/DRC. A simple model is shown which captures infection dynamics and allows for a concomitant increase in population size, *N*. The change in population size in Kinshasa, and the Democratic Republic of the Congo as a whole, may be observed from the plot of census data over time (Appendix E, Figure E4).

Chapter 11

Related publications

1. Lycett, S. J., M. J. Ward, F. I. Lewis, A. F. Y. Poon, S. L. K. Pond, *et al.* (2009). Detection of Mammalian Virulence Determinants in Highly Pathogenic Avian Influenza H5N1 Viruses: Multivariate Analysis of Published Data. *Journal of Virology* **83**(19): 9901-9910

2. Volz, E. M., S. L. K. Pond, M. J. Ward, A. J. L. Brown and S. D. W. Frost (2009). Phylodynamics of Infectious Disease Epidemics. *Genetics* **183**(4): 1421-1430

3. Volz, E. M., J. S. Koopman, M. J. Ward, A. J. Leigh Brown and S. D. W. Frost (2012). Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection. *PLoS Computational Biology* **8**(6): e1002552

JOURNAL OF VIROLOGY, Oct. 2009, p. 9901–9910 Vol. 83, No. 19
0022-538X/09/$08.00+0 doi:10.1128/JVI.00608-09

Detection of Mammalian Virulence Determinants in Highly Pathogenic Avian Influenza H5N1 Viruses: Multivariate Analysis of Published Data[▽][†]

S. J. Lycett,[1]* M. J. Ward,[1] F. I. Lewis,[1]‡ A. F. Y. Poon,[2]
S. L. Kosakovsky Pond,[2] and A. J. Leigh Brown[1]

Institute of Evolutionary Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom,[1] and University of California at San Diego, 150 West Washington Street, Suite 100, San Diego, California 92103[2]

Received 24 March 2009/Accepted 10 July 2009

Highly pathogenic avian influenza (HPAI) virus H5N1 infects water and land fowl and can infect and cause mortality in mammals, including humans. However, HPAI H5N1 strains are not equally virulent in mammals, and some strains have been shown to cause only mild symptoms in experimental infections. Since most experimental studies of the basis of virulence in mammals have been small in scale, we undertook a meta-analysis of available experimental studies and used Bayesian graphical models (BGM) to increase the power of inference. We applied text-mining techniques to identify 27 individual studies that experimentally determined pathogenicity in HPAI H5N1 strains comprising 69 complete genome sequences. Amino acid sequence data in all 11 genes were coded as binary data for the presence or absence of mutations related to virulence in mammals or nonconsensus residues. Sites previously implicated as virulence determinants were examined for association with virulence in mammals in this data set, and the sites with the most significant association were selected for further BGM analysis. The analyses show that virulence in mammals is a complex genetic trait directly influenced by mutations in polymerase basic 1 (PB1) and PB2, nonstructural 1 (NS1), and hemagglutinin (HA) genes. Several intra- and intersegment correlations were also found, and we postulate that there may be two separate virulence mechanisms involving particular combinations of polymerase and NS1 mutations or of NS1 and HA mutations.

H5N1 avian influenza initially garnered widespread attention following the first human cases of infection recorded in Hong Kong in 1997, after an outbreak in chickens (15, 72, 74), which itself followed an earlier infection in water fowl (21, 84). Since then the virus has spread across Asia to Russia, the Middle East, Europe, and Africa, causing deaths in wild aquatic birds (13, 58, 71, 81), domestic ducks and chickens (29, 39, 43), dogs and cats (2, 35, 36, 55, 76), and 254 of 405 confirmed human cases as of 5 February 2009 according to the World Health Organization (http://www.who.int/csr/disease/avian_influenza/).

Avian influenza virus strains are classified as highly pathogenic or of low pathogenicity on the basis of their virulence in chickens—highly pathogenic avian influenza (HPAI) virus can cause significant mortality (75 to 100%) in unvaccinated flocks (see reference 1), whereas low-pathogenicity strains cause only mild or no symptoms. HPAI H5N1 strains possess multiple basic amino acids in the hemagglutinin (HA) surface glycoprotein (27) cleavage site, the characteristic of all HPAI (62, 63). A variety of substrains of HPAI H5N1 have appeared during the last 10 years (82) and have probably become endemic in both wild birds and poultry (12, 43, 58). All strains of HPAI H5N1 are highly pathogenic in chickens, but their virulence varies in ducks (29) and mammals (see for example, references 11, 20, and 34).

The influenza A virus has an eight-segment RNA genome encoding 11 proteins, and the effects of point mutations or allelic combinations on the virulence of HPAI H5N1 isolates in mammals have been measured in mammals by reverse genetics and reassortment studies. Virulence in mammals is thought to be polygenic (10, 20), and several experimental studies have shown that mutations in the genes encoding internal proteins are important virulence determinants. For example, E627K in PB2 was shown to be important for replication in mammalian cells (73) and in the difference in virulence of two H5N1 Hong Kong 1997 outbreak isolates (24), whereas D92E in NS1 was found to increase resistance to tumor necrosis factor alpha and gamma interferon host responses for another H5N1 Hong Kong 1997 strain in vitro and in vivo in one study in swine (67).

Given the diversity of H5N1 viruses, their high rates of mutation and reassortment (14, 21, 22, 68, 77), and the polygenic nature of their virulence in mammals, general conclusions about virulence factors cannot be drawn with confidence from individual studies of small numbers of closely related sequences. However, systematically amalgamating the results from several studies may allow a more comprehensive picture to emerge. Such integration of results from many experiments and assays to provide a more complete understanding of gene-protein interactions has been carried out in *Saccharomyces*

* Corresponding author. Mailing address: Ashworth Laboratories, Rm. 65, Institute of Evolutionary Biology, University of Edinburgh, Kings Buildings, West Mains Road, Edinburgh EH9 3JT, United Kingdom. Phone: 44 131 650 8683. Fax: 44 131 650 5456. E-mail: samantha.lycett@ed.ac.uk.

† Supplemental material for this article may be found at http://jvi.asm.org/.

‡ Present address: Epidemiology Research Unit, Drummondhill, Stratherrick Road, Scottish Agricultural College, Inverness IV2 4JZ, United Kingdom.

▽ Published ahead of print on 22 July 2009.

cerevisiae (32, 40, 65). Using multiple data sources reduces the number of false-positive results and can reveal new patterns of association (32, 40), provided data quality and content issues are overcome (see for example, reference 56). Furthermore, while assembly of a few large data sets into a form suitable for meta-analysis can be done manually, experimental results from many small individual studies can also be captured by text mining published abstracts from online databases such as PubMed at the National Center for Biotechnology Information (NCBI) (51, 69).

The aim of the present study was to find a statistical model to describe the amino acid sites important for virulence of HPAI H5N1 in mammals and to indicate possible interactions between the sites. We have text mined PubMed abstracts for H5N1 isolates with experimentally determined virulence in mammals and downloaded the corresponding full genome sequences from the NCBI Influenza Virus Resource (4). We then fitted a Bayesian graphical model (BGM) to capture the dependencies among mutations and between individual mutations and the mammalian virulence phenotype. BGMs represent the probabilistic dependencies between random variables as a network without self-references, i.e., a directed acyclic graph (37, 59). They provide a convenient description of multivariate data because each link encodes a direct probabilistic dependency between sampled variables (e.g., $A - C$), rather than just a correlation (e.g., $A - C$, because $A - B$ and $B - C$). BGMs have been used in several biological contexts: to infer gene expression networks from microarray data (18), to find the genes involved in a complex trait using single nucleotide polymorphism data (66), and to find coevolving sites in a rapidly variable region of a human immunodeficiency virus (HIV) gene (61). In the present study the nodes (random variables) in the network represent variable amino acid sites or the phenotype: links between nodes represent direct dependencies, i.e., links between amino acid sites represent putative functional interactions, and links between amino acid sites and the phenotype of "virulence in mammals" represent direct association of specific mutations and virulence.

MATERIALS AND METHODS

Text mining. A list of H5N1 HPAI isolates with experimentally determined virulence in mammals was obtained by using a text-mining approach, followed by manual curation. Initially, XML (Extensible Markup Language) records of all abstracts containing "H5N1" and "virulence" keywords were obtained from PubMed using Eutils queries (http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) generated with R scripts (http://www.r-project.org). Each of the 542 abstracts obtained on 5 March 08 were searched for isolate names (such as A/Vietnam/1203/2004) using a "/(words)/(words)/" regular expression (i.e., at least three "/"s). This yielded 185 isolate names in 97 abstracts. Next, a list was compiled (using R scripts and manual editing) of unique isolate names, taking into account synonyms (e.g., dk = duck) and recording the PubMed identifiers (PMIDs) of papers that had mentioned the strain. Of the 94 unique isolate names found, 64 were for H5N1 isolates with measured virulence in mammals and full genome sequences deposited in the NCBI Flu database (4) (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html).

The abstracts corresponding to the 64 H5N1 isolates were examined, and detailed results from mice or ferret studies were manually extracted from the corresponding studies. In some studies additional strains not mentioned in the abstract had also been studied, and these strains were also added to the original list, bringing the total to 69 strains from 27 individual studies. Table SA1 in the supplemental material contains the references to the studies and the numbers of strains from each study in the final data set. Each strain was manually classified as either virulent or nonvirulent in mammals based upon the amalgamated experimental evidence. Table SA2 in the supplemental material contains the assigned classification and experimental evidence per strain.

Data preprocessing. Full genome nucleotide sequences of the 69 isolates with experimental evidence for (or against) virulence in mammals were downloaded from the NCBI Flu Database. The individual segments were initially aligned by using CLUSTAL W in BioEdit, and then a final manual codon alignment was performed in MEGA 4.0. The PB1-F2, M2, and NS2 coding sequences were manually extracted from the aligned nucleotide files (not all strains were annotated for these alternative coding sequences), and all 11 nucleotide coding sequences were translated to protein sequences in MEGA 4.0.

To analyze the association of particular residues at amino acid sites and virulence using BGMs, the protein sequences were converted into a binary matrix where each row of the matrix represented an isolate genome (11 proteins concatenated), and columns represented amino acid sites. An additional column for the virulence phenotype was added. Zeros represented consensus amino acid residues, ones represented any mutant (or the virulent phenotype). In particular, the following procedure was used to determine the binary coding at each variable site (via a custom R script). (i) If there are only two types of amino acids present at a site (A or B), a 2×2 table was calculated, counting the number of A residues in the virulent sequences, V(A); the number of A's in nonvirulent sequences, N(A); the number of B's in virulent sequences, V(B); and the number of B's in nonvirulent sequences, N(B). (ii) To determine which to code as 1, A or B, the determinant of the 2×2 table was calculated as $V(A)N(B) - V(B)N(A)$. If the determinant is positive, then amino acids of type A are coded as 1's and the B's are coded as 0's (and vice versa for negative determinants). Alternatively, if there is evidence in the literature to indicate that a particular residue is related to virulence (or nonvirulence), then this residue is coded as 1 (or 0). (iii) If there are more than two types of residue at a site, then step 2 is repeated for all combinations of amino acids partitioned into two groups, and the combination with the smallest P value from the Fisher exact test on the 2×2 table is chosen.

The coding scheme used, corresponding 2×2 table entries and P value from the Fisher exact test for sites of interest are reported in Table 1 (see Table SA3 in the supplemental material for information on all of the sites examined).

Multivariate analysis of variable amino acids. (i) BGMs. BGMs were inferred by using HyPhy (60, 61) from sequence and phenotype data coded to a binary matrix. HyPhy implements the methods of Cooper and Herskovits (16) and Friedman and Koller (19) for BGM inference as follows. The probability for a particular graph structure given the data can be calculated by using the K2 scoring metric (16). The K2 metric score for one node is a function of the number of discrete states of the node and its parents (e.g., how many $B = 1$ when $A = 1$, etc.), and the total score for the graph is the product of the individual node scores. When the number of nodes is nontrivial (typically greater than 5), evaluating all possible graph structures becomes prohibitive so the Monte Carlo Markov chain (MCMC) method is used over families of ordered nodes to find a set of probable graph structures to describe the data (19). The output of the BGM inference is a consensus graph, where each link is assigned a probability of existing according to how often it appears in the sampled MCMC results (i.e., the marginal posterior probability − the expectation of posterior probabilities weighted by the likelihoods of the node orderings in which the given link appears). Links with probability greater than or equal to 0.5 in the consensus graph are used to create the final network.

For analysis of the 69 sequence data set (with 10 nodes), the HyPhy BGM MCMC was run with 10^6 steps (after a burn-in of 10^5 steps) and sampled every 1,000 steps. These parameters were chosen so that the sampled likelihood values were converged to a stationary distribution, and the individual samples were independent (the autocorrelation of the sampled likelihood values was calculated in every run, and the values of the first lags were not significantly different from zero). The maximum number of parents allowed per node was increased until there was no significant change in network structure (i.e., no new links where $P \geq 0.5$ occurred), and a maximum of five parents per node was found to be sufficient for the present data set.

(ii) Model validation. The significance of the links in the BGM inferred from the 69 × 10 binary data matrix was tested in two ways.

(a) The inferred BGM was compared to a set of null model BGMs. Null models were generated by inferring BGMs from nonparametrically bootstrapped data in which one site (or phenotype) was randomly permuted (i.e., a fixed column in a matrix of observation was resampled without replacement) at a time (leaving the other sites unchanged). Ten randomizations were performed per site (i.e., one-hundred randomizations for the ten sites). Randomizing each site (or the phenotype) independently of the others and inferring the BGM gives the background probability of the links between that site and the others due to random noise alone, conditional upon the joint probabilistic structure of the remaining data. The unpermuted BGM link probabilities from the site of interest

TABLE 1. Amino acid site coding and *P* value for association with virulence using the Fisher exact test[a]

Site	Residue(s)		*P*	Selected	Reference(s)	Additional detail(s)
	Virulent	Nonvirulent				
PB2-318	K	R	0.0027	*	10	
PB2-355	K	R, Q	0.0003	(PB1-317)	10, 34, 41	
PB2-627	K	E	0.0034	*	10, 11, 24, 41	
PB2-627 or 701	K, N	E, D	0.0003	(*)	17, 38, 70	
PB1-317	I	M, V	3.8E-05	*	10, 34, 41	
PA-127	I	V	0.0007	(PB2-318)	75	
PA-336	M	L	0.0006	(PB2-318)	75	
HA-102 (86E)	V	A, I, P, S, T	0.0092	*	83	
HA-140 (124B)	S	N, D	0.0431		83	
HA-154 (138A)	L, N	Q, H, I	0.0391		83	
HA-172 (156 glyco)	T, S	A	0.0010	*	8, 10, 83	
HA-228 (212D)	E, R	K	0.0194	*	83	
HA-279 (263E)	T	A	1.8E-05	*	83	
NA-49:72	Any deletion	No deletion	0.0003	*	5, 52	Stalk deletion
NS1-42	S	A, P	0.0244		33	P attenuates pathogenicity
NS1-92	E	D	0.0010	*	41, 67	
NS1-92/97	E	D	0.0504		48	D92E (no deletion) or D97E (with deletion)
NS1-127	N	T, D, R, V, A	0.0387		53	123-127 PKR region
NS1-189	N	D, G	0.0006	(PB2-318)	75	
NS1-195	T, Y	S	0.0294	(HA-102)	7	
NS1-228	P	S	0.0027	(NS1-92)	31, 57	PDZ binding domain
NS2-31	I	M	0.0006	(PB2-318)	75	
NS2-56	Y	H, L	0.0006	(PB2-318)	75	

[a] For each sequence, sites containing the "virulent" residues (column 2) were coded as ones or zeros (column 3). The numbers of sequences classified as virulent (nonvirulent) with virulent (nonvirulent) residues were calculated, and the *P* value from the Fisher exact test (uncorrected) is displayed in column 4. The nine sites with the lowest *P* values selected for further analysis are indicated by an asterisk in column 5; sites with the same (or similar) pattern of mutations to selected sites are also indicated in column 5. All site numbering is from the first methionine in the coding sequences. The numbers in parentheses for HA correspond to the mature H5 site numbers, as in reference 83, and the letters correspond to the canonical antigenic site, or a glycosylation site (glyco). Abbreviations: PB, polymerase basic; NP, nucleoprotein; NA, neuraminidase; M, matrix protein; NS, nonstructural protein.

to or from the other sites were normalized by the mean and standard deviation of the 10 permuted results to give a standard score per link.

(b) Cross-validation of structure was applied. To test the robustness of the inferred BGM and mitigate any possible effects from isolate misclassification or sequencing errors, 10 BGMs were calculated and averaged over different subsets of the data. Each subset of data consisted of 65 sequences (i.e., training data), and the 4 remaining sequences were kept as test sets. Since a different group of strains (particularly the Z constellation) began to replace the early strains from 2002 onward (11, 22, 43), different patterns of mutations may occur after 2002; hence, four sequences were excluded in each run, one from each of the pre-2002 (early) virulent, 2002+ (late) virulent, pre-2002 nonvirulent, and 2002+ nonvirulent sequences.

(iii) Conditional probability tables. Conditional probability tables describing the probability that a strain is virulent (nonvirulent) given the presence or absence of mutations at individual sites (identified in the BGM as directly influencing virulence), or a combination of those sites, were calculated separately for each of the 10 cross-validation training data sets using the observed frequency of mutations. Certain combinations of mutations in the influencing sites were not observed because of the correlation between the sites (e.g., there were no sequences in the 69 sequence set with PB1-317 = 1 and NS1-92 = 0 or vice versa). For (hypothetical) strains with these unobserved combinations, we estimated the probability of virulence from the probability of virulence of the individual sites, assuming that the sites were independent as in equation 1 below.

$$P(Vir = 1|a_1a_2a_3) = 1 - \prod_{i=1}^{3}[1 - p_i(a_i)] \qquad (1)$$

where $P(Vir = 1|a_1a_2a_3)$ is the probability of virulence for a sequence with the particular combination $a_1a_2a_3$ of amino acids, and $p_i(a_i)$ is the probability of virulence when node *i* has amino acid *i*.

The performance of the model was examined by predicting the virulence of the sequences in each of the 10 training and test data set pairs using the conditional probability table derived from each training set, respectively. The numbers of true positives (*TP*; virulent sequences predicted to be virulent by the model), false negatives (*FN*; virulent sequences predicted as nonvirulent), false positives (*FP*; nonvirulent sequences predicted as virulent), and true negatives (*TN*; nonvirulent sequences predicted as nonvirulent) were calculated along with the misclassification ($FP + FN$), sensitivity [$TP/(TP + FN)$], and specificity [$TN/(FP + TN)$] values.

RESULTS

Isolates with experimentally determined virulence in mammals. A total of 69 H5N1 full-genome isolates from 27 studies of virulence in mammals identified using a text-mining approach, followed by manual curation, were used in this analysis (see Materials and Methods and Table SA1 in the supplemental material). The most frequently measured strain—A/Vietnam/1203/04—was the subject of seven virulence studies. Five studies (11, 20, 34, 47, 50) contained measurements for at least five sequences. Strains were classified as virulent (or nonvirulent) if infection studies showed a high mortality with a low dose; sequences with 50% lethal dose of $\leq 10^3$ 50% egg infectious doses were classified as virulent in mammals. Even though the virulence results were obtained in different animals and using different protocols, the results were generally concordant: in 75% of cases (52 sequences), strains could be classified clearly either as highly pathogenic in mice or ferrets or only caused mild or no symptoms. Furthermore, three studies measured virulence in both mice and ferrets (28, 50, 64) in a total of five strains. In these cases, the pathogenicity in ferrets correlated with that in mice (highly pathogenic strains in ferrets were highly pathogenic in mice, low-pathogenicity strains in ferrets showed low pathogenicity in mice). Although differences in the virulence of two H5N1 strains in mice and ferrets

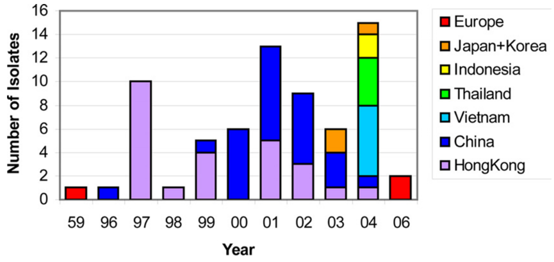


FIG. 1. Distribution of full-genome isolates in text-mined data across isolation year and region. The total number of isolates in the data set per year is indicated by the height of the bars, and the distribution by region is represented by the bar colors. The data covers three major outbreaks: Hong Kong in 1997, Hong Kong/China in 2001, Vietnam (plus others) in 2004 over a period of at least 10 years.

at high dose have been reported (86), we did not find significant differences in the overall virulence of the strains in the present data set when we used our binary classification scheme. Details of the strain classification and associated experimental evidence can be found in Table SA2 in the supplemental material.

The final full data set consisted of 69 full genome sequences: 35 from water fowl, 17 from land fowl, and 17 from humans. Totals of 20% (7) of the water fowl isolates, 24% (4) of the land fowl isolates, and 65% (11) of the human sequences were classified as virulent in mammals, making a total of 22 virulent sequences. The distribution of the strains across isolation year and place is shown in Fig. 1: this data set covers several outbreaks, including the 1997 Hong Kong, 2000 to 2002 China, and 2004 Viet Nam epidemics.

Site association with virulence. Across the whole genome, 1,235 of 4,582 (~27%) amino acid sites were variable (≥1 alternative residue in the 69 sequences) with 969 of 1,235 (~78%) of these variable sites containing only 2 amino acid residues. Furthermore, in the remaining 266 (22%) of variable sites that coded for three or more alternative amino acids, the average frequency of the next most abundant residue was only 4%. To facilitate analysis of the relationship between mutations and virulence in mammals, the sequence data were coded to a binary matrix, which therefore retained almost all of the information contained in the sequence alignment.

All 1,235 polymorphic amino acid sites in the protein sequences of the 69 strains were examined, and the distribution of the alternative residues among strains classified as virulent or nonvirulent was recorded. A total of 227 amino acid sites (19% of the variable sites) across the entire genome were found to have an association with virulence at $P \leq 0.05$ (Fisher exact test, uncorrected). However, there was a high degree of association among sites, with mutations at 129 sites having an identical distribution across the strains to at least one other site. Of the 98 sites with a distribution that was unique in the data set, only 53 differed by more than one mutation from any other site. Of the 98 sites with a unique distribution, 12 have been reported in the literature to have functional significance for virulence (see Table SA3 in the supplemental material).

We next examined all amino acid sites or genomic features (e.g., neuraminidase [NA] stalk deletion; PDZ binding domain in NS1) previously reported in the literature as having functional significance for virulence in H5N1 and found evidence for 70 sites/features. Their distribution among virulent and nonvirulent strains was tested as described above: the 23 that were apparently associated with virulence in this data set are shown in Table 1 (Fisher exact test $P \leq 0.05$, uncorrected; see Table SA3 in the supplemental material for details of all 70 sites). All of the strains in this data set, including nonvirulent ones, contained the HA-226(Q) and HA-228(G) (H3 numbering) residues for the avian sialic acid receptor binding site (23, 30) and the NS1-149(A) virulence determinant in chickens (45); on the other hand, only one strain contained the pathogenicity-associated proline at amino acid site 42 in NS1 (33).

Of the 23 sites, three in PB2 (sites 318, 355, and 627), one in PB1 (site 317) and one in NS1 (site 92) which were apparently associated with virulence ($P < 0.01$; Table 1) had previously been claimed to have a role in virulence determination (see the Discussion). In addition, the presence of PB2-701N has been suggested as an alternative virulence or adaptation marker in mammals in place of PB2-627K (17, 38, 70), and combining the PB2-627 and PB2-701 sites resulted in a more significant P value ($P = 0.0003$) for association with virulence than PB2-627 ($P = 0.003$) or PB2-701 ($P = 0.08$) alone. A further five sites in polymerase acidic (PA), and nonstructural genes (NS1 and NS2) were also nominally significant; however, these sites had very similar distributions across strains to PB2-318. Two HA antigenic sites—HA-102(mature H5 site 86, part of canonical antigenic site E) and HA-279(263E) and the glycosylation site HA-172(156)—had a possible association with virulence ($P < 0.01$), and deletions in the NA stalk region appeared to be quite strongly associated: 21 of the 22 virulent sequences showed stalk deletion compared to about half of the nonvirulent sequences ($P < 0.001$).

Site selection for BGM. The uncorrected associations are not themselves evidence of causal dependencies but assist with the selection of sites to be analyzed in the multivariate model. Sites with the same (or only one mutation different) pattern across the sequences were combined since the individual contribution of these sites to virulence would not be distinguishable from this data set. The sites with common patterns are shown in Table 2. Of the 23 sites and/or features with Fisher exact test P values of ≤0.05, only 13 had distinct patterns. In order to reduce the possibility of inferring false links, the total number of variables included in the BGM analysis was restricted as indicated by simulation studies based on sample sizes of 60 to 70 strains (see the discussion of simulation analysis in the supplemental material). Consequently, the nine

TABLE 2. Amino acid sites with similar patterns of mutations[a]

Pattern	Sites (virulent|nonvirulent)
1	PB2-318(K|R), PA-127(I|V), PA-336(M|L), NS1-189(N|DG), NS2-31(I|M), NS2-56(YL|H)
2	PB2-355(K|RQ), PB1-317(I|MV)
3	PB2-675(L|I), PB2-683(T|A), PB1-198(K|R), NA-39(QH|K), NA-223(I|T)
4	HA-102(V|AIPST), NS1-195(TY|S)
5	NS1-92(E|D), NS1-228(P|S)
6	M2-64(SAF|P), M2-69(P|L)

[a] The distributions of mutations across strains for sites identified in the literature (see Table 1 and Table SA3 in the supplemental material) were examined. These groups of sites have more than one mutation and identical or nearly identical distributions (≤1 mutation difference).

TABLE 3. Sites selected for BGM[a]

Site (virulent|nonvirulent)	Comment(s)
Vir	"Virulent-in-mammals" phenotype
PB2-318(K|R) ++	Combined pattern 1, see Table 2
PB1-317(I|MV)/PB2-355(K|RQ)	Combined pattern 2, see Table 2
PB2-627(K|E)	
HA-102(V|AIPST)/NS1-195(TY|S)	Combined pattern 4, see Table 2
HA-172(TS|A)	
HA-228(ER|K)	
HA-279(T|A)	
NA-del	NA stalk deletion ("1") or no deletion ("0")
NS1-92(E|D)/NS1-228(P|S)	Combined pattern 5, see Table 2

[a] The sites included in the BGM analysis were the nine sites with the most significant P values for association with virulence in mammals, and the "virulent-in-mammals" phenotype. Sites with the same distribution of mutations across the strains are analyzed together as indicated.

uniquely patterned sites with the most significant P values for association with virulence in mammals (ranging from $P < 10^{-4}$ to $P = 0.019$), together with the virulence phenotype, were included in the multivariate analysis (Table 3). The selected sites were: PB2-318, PB2-627, PB1-317 (≡ PB2-355), HA-102 (≡ NS1-195), HA-172, HA-228, HA-279, NA stalk deletion, and NS1-92 (≡ NS1-228) (the final binary data used can be found in Table SA4 in the supplemental material). Since there is evidence in the literature to suggest that PB2-701N may compensate for the lack of PB2-627K in the adaptation of the virus to mammals (17, 38, 70), we also investigated the effect of using the combined PB2-627 and PB2-701 site data in place of PB2-627 alone.

BGM results. BGMs were inferred from the binary data of the selected sites using: (i) all 69 sequences; (ii) 100 cases of permuted binary data (null models) and; (iii) 10 cases of leaving 4 sequences out each time (10× cross validation, see the discussion of model validation in Materials and Methods). Figure 2 shows the inferred network structure for significant links, together with the average link probability from the 10× cross validation and standard score from the comparison to the null model networks. The model confirms that virulence in mammals is a complex genetic trait affected directly by mutations in the polymerase, NS1, and HA genes.

Very strong correlations among certain sites leave some ambiguity over the precise nature of the molecular changes since these sites could not be analyzed separately (Table 2). Specifically, three distinct mutation patterns, representing two sites each—PB1-317/PB2-355, NS1-92/NS1-228, and HA-102/NS1-195—were directly associated with virulence in mammals in the model. We did not detect a significant direct association of PB2-627 alone with virulence in mammals in the model; however, a strong direct association, with a probability of ≥0.8 (standard score = 3 to 4) was detected for the combined sites PB2-627 and PB2-701 (data not shown).

Sites within the polymerase genes were strongly associated with each other in the model: the pattern 1 sites (including PB2-318, PA-127, and PA-336) were linked to PB1-317/PB2-355 pattern 2 sites with a probability of ≥0.9 (standard score = 9 to 10). Strong intrasegment associations were also found in both HA and NS, particularly between antigenic sites HA-102 and HA-279, HA-102 and the variable glycosylation site HA-172, and NS1-92 and NS1-228 (pattern 5) and between NS1-189, NS2-31, and NS2-56 (pattern 1). The most significant intersegment association was between PB1-317/PB2-355 (pattern 2) and NS1-92/228 (pattern 5). NS1 sites 92, 195, and 228 were also strongly associated with HA-279 (antigenic site), and NS1-189 had the same pattern as the PB2-318, PA, and NS2 sites (pattern 1). Additionally, the previously reported interaction between the HA-172(156) glycosylation site and the NA stalk deletion (3) was detected in the BGM with a probability of ≥0.8 (standard score = 9 to 10).

The model was tested by using cross-validation (see Materials and Methods). Table 4 and Table 5 show the conditional probabilities for virulence in mammals averaged over each training data case from the cross-validation data sets, calculated per single influencing node (Table 4) and per combination of influencing sites (Table 5). Where a particular combination of amino acids has not occurred in this data set (due to the small data set size and interactions between the influencing sites), the probability of virulence in Table 5 was estimated by combining the individual probabilities in Table 4, assuming independence among sites (see the discussion of conditional probability tables in Materials and Methods).

Using the conditional probability tables from the 10 training data examples, virulence predictions were made for each test data example (these were composed of the four excluded sequences from the training data: one of each early-

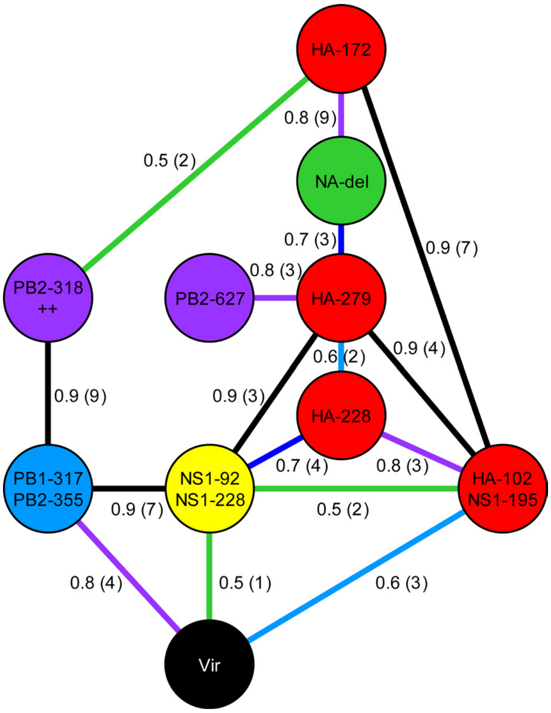


FIG. 2. Inferred BGM for virulence in mammals. The indicated link probabilities from 10× cross validation and z-score from comparison to null models (in parentheses) are rounded down, e.g., "0.8 (3)" means a link probability of 0.8 to 0.9 and a z-score of 3 to 4. Note that some nodes represent more than one site, as listed in Table 2. The sites directly associated with virulence in mammals are PB1-317/PB2-355, NS1-92/NS1-228, and HA-102/NS1-19.

TABLE 4. Conditional probabilities for “virulent-in-mammals” sequences given individual sites[a]

Node	Sites	Vir-AA	Non-Vir AA	P(virulence|Vir-AA)	P(virulence|non-Vir AA)
1	PB1-317/PB2-355	I/K	MV/RQ	1.00	0.22
2	NS1-92/NS1-228	E/P	D/S	0.73	0.22
3	HA-102/NS1-195	V/TY	AIPST/S	0.69	0.24

[a] Given that a sequence has a “virulent” type amino acid (Vir-AA) at a site (as indicated in column 3), the probability of that sequence having a “virulent” classification is indicated in column 5. If the sequence has a “nonvirulent” amino acid (Non-Vir AA) at a site (column 4), the probability of that sequence still having a “virulent” classification is indicated in column 6. For example, if the sequence contains PB1-317I (and/or PB2-355K), the probability of that sequence being virulent is 1; if the sequence does not have these mutations, then the probability of virulence is 0.22. All probabilities were calculated by averaging over the 10 training data examples and assuming independence between sites (rows).

TABLE 6. Average prediction performance over cross-validation training and test data sets[a]

Data set	Avg performance (%)	
	Training	Test
TP	13.6 (20.9)	1.4 (35)
FN	6.4 (9.8)	0.6 (15)
FP	2.9 (4.5)	0.1 (2.5)
TN	42.1 (64.7)	1.9 (47.5)
MC	9.3 (14.3)	0.7 (17.5)

[a] Each sequence in the training and test data sets was classified as virulent or nonvirulent according to the combination of amino acids at PB1-317/PB2-355, NS1-92/NS1-228, and HA-102/NS1-195 as in Table 5. The average numbers (or percentages) of true positives (TP; virulent sequences classified as virulent), false negatives (FN), false positives (FP), and true negatives (TN) from the 10 training and test data sets, and the misclassification (MC) performance metrics are indicated. The sensitivity and specificity values were 68% and 94%, respectively, for the training data sets and 70% and 95%, respectively, for the test data sets.

virulent, late-virulent, early-nonvirulent, and late-nonvirulent sequences, where early refers to a pre-2002 occurrence; see the discussion of model validation in Materials and Methods for details). The average prediction performance over the cross validation training and test data sets was calculated (Table 6); for the test data, the sensitivity was 70%, and the specificity was 95%.

DISCUSSION

This study combines results from previous experimental studies based on strains isolated during 3 major HPAI H5N1 outbreaks in order to examine the genetic basis of virulence in mammals. Extensive literature mining, followed by rigorous quality control, identified 69 complete genome sequences with experimentally determined virulence in mammals for use in this analysis, a threefold increase from the number of sequences analyzed in the largest experimental study to date. We have fitted for the first time a rigorous statistical model relating mutations in multiple influenza viral segments to virulence in mammals. Many authors have previously described virulence in terms of a complex trait (10, 11, 20): the model presented here confirms that virulence in mammals is affected by mutations in at least four segments: PB1, PB2, HA, and NS1.

TABLE 5. Conditional probabilities for “virulent in mammals” sequences for combinations of sites[a]

Node[b]			P(virulence|state)[c]	Virulence prediction ($P \geq 0.5$)
1	2	3		
0	0	0	0.09	0
0	0	1	0.70	1
0	1	0	0.25	0
0	1	1	0.94*	1
1	0	0	1*	1
1	0	1	1*	1
1	1	0	1	1
1	1	1	1*	1

[a] A sequence with the combination of mutations [“virulent” (1)/“nonvirulent” (0) amino acids] as indicated in columns 1 to 3 has a probability of virulence in mammals as indicated in column 4. If the probability of virulence in mammals is >0.5, then the sequence is classified as ‘virulent’.

[b] Node 1, PB1-317/PB2-355; node 2, NS1-92/NS1-228; node 3, HA-102/NS1-195.

[c] The values marked with an asterisk indicate that this particular combination of mutations was not observed in the training data sets (or 69 sequence data set), so the probability of virulence in mammals was estimated from the individual values in Table 4 (see Materials and Methods). All probabilities were calculated by averaging over the 10 training data examples.

Univariate analyses. Three sites in PB2 (sites 318, 355, and 627), PB1-317 and NS1-92 that were apparently associated with virulence in the univariate studies had been proposed in earlier experimental studies (10, 11, 24, 34, 41, 50, 67). Recent experimental evidence indicates that a PB2-701N mutation may arise in mammals instead of PB2-627K (70); we found no isolates containing both mutations in the 69-sequence data set, and combining these two sites increases the association with virulence. A further five sites in PA, NS1, and NS2 identified in a study of the 1997 Hong Kong H5N1 human outbreak (75) were also nominally significant; however, these sites had very similar mutation patterns to PB2-318 and the “virulent” residues only appeared in the 1997 sequences. We did not detect any significant association of NS1 deletions alone with virulence in mammals in this data set (48). However, since NS1-97E is a conserved residue (in the aligned sequences), the 5-amino-acid NS1-deletion also leads to an E at position 92 in the protein (48), and NS1-92E (no deletion) or NS1-97E (with deletion) with virulence in mammals was marginally significant ($P \sim 0.05$). The association of NS1-92E alone (aligned sequences) with virulence in mammals was more significant in this data set ($P < 0.001$). In addition, no significant associations between virulence in mammals and the Src homology 3 domain binding motifs or PDZ binding domains containing the amino acid motif ESEV or EPEV in NS1 were found (26, 31, 57). However, proline at NS1-228 (part of the PDZ binding domain) was associated with both virulence in mammals ($P < 0.003$) and strongly with E at NS1-92 ($P \ll 10^{-6}$).

Multivariate BGM models. The nine sites with the strongest individual associations with virulence were included in the BGM analysis, and three of them (PB1-317/PB2-355, NS1-92/NS1-228 and HA-102/NS1-195) were found to have a direct association with virulence.

The mutation patterns at PB1-317 and PB2-355 were very similar to each other in the 69-sequence set and directly associated with virulence in mammals. These mutations were first identified by Katz et al. (34) as correlated with virulence in mice in the 1997 Hong Kong outbreak. In this 69-sequence data set, the PB1 (M,V)317I and PB2 (R,Q)355K “virulent” mutations only appear in sequences from the 1997 Hong Kong outbreak. In an alignment of 579 full-genome sequences (representing all H5N1 unique sequences deposited in the NCBI Influenza Virus Resource as of September 2008), 44 sequences contained one and 12 contained both of the mutations. Apart from the 1997 Hong Kong outbreak, PB1-317I and/or PB2-355K mutations were also present in sequences from birds in China and Viet Nam in 2005 and 2006; a civet (Viet Nam 2005); and four human sequences (Indonesia 2006, China 2006 and 2007).

Two mutations in NS1, D92E and S228P, which were strongly associated in the data set were also directly associated with virulence in mammals in the BGM. The NS1 D92E change is thought to influence pathogenicity by making the virus more resistant to host interferon and tumor necrosis factor alpha responses (67), thus inducing a more severe cytokine response in the host (46). NS1-228 is one of the C-terminal residues and part of a region that could bind proteins containing PDZ domains. NS1 proteins with avian ESEV or EPEV PDZ domain ligand motifs were postulated to disrupt human cell pathways by binding to proteins with PDZ domains (57), and NS1 proteins with avian PDZ domain ligands have been shown to cause increased virulence in mice (31). A total of 51 of the 69 sequences examined here contained ESEV or EPEV at positions 227 to 230: no significant association of these motifs with mammalian virulence was found. However, the point mutation S228P was significantly associated with virulence in mammals. The NS1-92 and NS1-228 mutations remain correlated in the 579 H5N1 full genome sequence set. There are 17 sequences with both mutations, the majority were from the 1997 Hong Kong outbreak, but 3 were from isolated from water fowl in Viet Nam 2005. Five sequences (one from swine in China 2003 and four from birds in Viet Nam and Russia 2005) contained the S228P mutation only, but no sequences contained only the D92E mutation. The BGM inferred from the 69-sequence data set showed a strong link between PB1-317/PB2-355 and NS1-92/NS1-228, and this association still holds in the 579 full-genome data set ($P \ll 10^{-6}$ for the Fisher Exact test with the Bonferroni correction for multiple testing between any of the 2,312 variable sites, i.e., 2.67×10^{6} pairwise interactions). All 12 sequences with both PB1-317 and PB2-355 mutations also contained both NS1-92 and NS1-228 mutations. Hence, the model and wider results indicate that the particular polymerase and NS1 gene combination significantly affects HPAI H5N1 virulence in mammals and that a mutation pattern associated with virulence in mammals in the 1997 Hong Kong outbreak has reemerged among human H5N1 cases in Indonesia and China in 2006 and 2007.

The other mutations directly associated with virulence in the BGM are NS1-195T and HA-102V. These sites are not strongly linked in the model to the PB1, PB2, and NS1 sites discussed above and may represent a separate mechanism for enhanced virulence in mammals. NS1 sites S195, T197, and D92 are hydrogen bonded in the NS1 crystal structure (6) and make a cleft. It has been proposed that mutation D92E could alter the phosphorylation of this NS1 cleft and thus affect virulence (6, 44). The BGM results indicate that NS1-195 mutations also affect virulence, and it is postulated that this could be due to NS1-195 mutations changing the phosphorylation properties of NS1, although clearly more experimental work is required to investigate this as a possible mechanism. Interestingly, the pattern of HA-102(86) V residues were completely correlated with the NS1-195 T in the 69 sequence data set. HA-102(86) is part of antigenic site E (83), so it is conceivable that a mutation here could reduce any cross-immunity that mammals may have against the virus. The correlation between NS1-195T and HA-102V is highly significant in the 579 full-genome sequence set ($P \ll 10^{-6}$ for the Fisher Exact test with the Bonferroni correction), and there are 83 sequences with both mutations, recorded from 2004 onward, mostly in Thailand, Viet Nam, and Russia. HA-102 is a highly polymorphic site, with six different residues present in the 69-sequence set. Of these, V (present in 10 isolates) and A (present in 55 isolates) are associated with virulence and nonvirulence in mammals, respectively. The remaining four residues (I, P, S, and T) are present in one isolate each, making a virulence assignment on the basis of the 69 sequence data set alone difficult. In the 579-sequence data set, residues I, P, and S were rare and only present in 13 avian isolates in total. HA-102T, on the other hand, was present in 48 isolates: 44 of the 48 human isolates from Indonesia; a cat isolate from Indonesia, and 3 avian isolates from China. In the BGM, HA-102V was directly associated with virulence with a link probability of 0.6 to 0.7 (standard score = 3 to 4), and updating the binary coding scheme for HA-102 to include T as a "virulent" residue increases the link probability in the BGM to 0.8 to 0.9 (the standard score remains 3 to 4). Consequently, HA-102 V or T should be considered as a virulence marker.

At first sight, it is surprising that PB2-627 alone was not found to be directly associated to virulence in mammals because this mutation is thought to be involved in adaptation to mammalian hosts (73), has been previously reported as contributing to virulence in the 1997 Hong Kong outbreak (10, 24, 25), and was shown to be important for efficient respiratory tract growth in mammals (25). The sequence set used here spanned more than the well-studied 1997 Hong Kong outbreak and contained 6 PB2-627K virulent sequences (of 22 virulent sequences), and 1 PB2-627K nonvirulent sequence (of 47). The six virulent PB2-627K sequences were sampled in Hong Kong in 1997 (2), Thailand and Viet Nam in 2004 (3), and Germany in 2006 (1), whereas the single PB2-627K sequence classified as nonvirulent was from A/Turkey/65596/2006 (a human case in Turkey), which caused only mild symptoms in ferrets at a low dose (85). In one of the cross-validation data sets, the PB2-627K nonvirulent sequence was (randomly) excluded, and the resulting BGM had a PB2-627-virulence link probability of 0.6. However, using only sequences that were either very highly pathogenic or nonpathogenic (75% of cases, 52 sequences), no significant link was found between PB2-627K and virulence in mammals. On the other hand, since it has been suggested that a PB2-701D-to-N mutation may arise to compensate for a the lack of the "virulent" PB2-627K residue in mammals (17, 38, 70), we combined the PB2-627 and PB2-701 sites, repeated the analysis, and found a significant direct association between the combined sites and virulence in mammals using BGMs (link probability of ≥0.8, standard score = 3 to 4). Although the relationship between PB2-627K alone and virulence in mammals was not found to be significant in the 69 sequence data set, these results suggest that PB2-627K and PB2-701N might be directly associated with virulence, but more isolates with these mutations and experimentally measured virulence in mammals would be required to detect the associations unambiguously. In addition, more isolates would allow the putative negative association between PB2-627K and PB2-701N to be evaluated.

Even though the uncorrected *P* value from the Fisher exact test for the association of NA stalk deletion and virulence in mammals was ~0.0003, the BGM results did not show a direct association. However, strong direct associations of NA stalk deletion with HA-172(156, glycosylation site) and HA-279 (263, antigenic site) were inferred. HA and NA have complementary functions in viral replication (see Wagner et al. [78] for a review), and various studies have indicated that the effects of mutations in one can be compensated by mutations in the other (52, 54, 79); consequently, some HA-NA associations are to be expected. In particular, the NA stalk deletion–HA-172 glycosylation intersegment association (3) was confirmed by the BGM in the present study. Both the association of HA-172 and NA stalk deletion and the HA-279 and NA stalk deletion were significant in the 579-genome sequence set (Bonferroni-corrected Fisher exact test result of $P < 10^{-6}$ for each). Neither the BGM nor the large sequence set provided any evidence that HA-172 and HA-279 were associated with each other (the corrected *P* value for the Fisher exact test on the large sequence set was 1), so we deduce that the NA stalk deletion is independently associated with both HA mutations, indicating a functional interaction between HA and NA.

An underlying issue with analysis of genetic determinants in influenza is the general lack of recombination within segments and consequent strong associations of amino acid variants over evolutionary time. Among the 228 sites with $P \leq 0.05$ for association with virulence in the 69-sequence data set, the high degree of correlation in patterns of mutation is largely due to identity by descent of the viral genome segments. Thus, 34 sites have most or all of their "virulent" residues in the 11 sequences isolated in the 1997 Hong Kong outbreak, 15 sites on segment 8 have mostly "virulent" residues except in the Goose/Guangdong/96 and 1999-2001 NS allele B sequences (9, 21, 49), and 7 sites have most or all of their "virulent" residues in the 10 sequences from the 2004 Vietnam/Thailand outbreak.

To mitigate the potential bias in the results introduced by nonindependence, an alternative approach would be to code each sequence for the presence or absence of mutations compared to its immediate inferred ancestral sequence (using a phylogenetic tree and model of sequence evolution). The change in the viral phenotype from the ancestral sequences would also have to be inferred. However, although this approach was successfully used to find mutations in the HIV envelope protein (60, 61), it cannot be applied to a data set where each segment has its own evolutionary history (22, 77, 80) since reassortment of the internal genes with other subtypes means that no single tree is applicable to all eight segments. Recognizing the issue, we reduced potential biases incurred due to the relationships between the sequences by using a relatively wide range of isolates (spanning more than 10 years) and selection of sites with dissimilar mutation patterns across the strains.

We used several techniques to reduce other potential sources of error in the model. First, backed by simulation studies investigating the number of false-positive associations as a function of data set size and number of sites, we limited the number of sites in our analysis to those with prior evidence from the literature and the strongest univariate associations to virulence in mammals. Second, we used cross-validation to reduce the effects of sequence misclassification (and sequence errors): the model presented here is an average over 10 subsamples of the data. Third, we used permutation tests to validate the significance of the associations against the null model of association by chance.

Using the model, we estimated the probability of virulence in mammals for strains containing different combinations of influencing sites. To evaluate the prediction performance of the model, we predicted each sequence in the cross-validation test data sets as virulent or nonvirulent on the basis of the combination of mutations it contained and compared the results to the original values. The high specificity (95%) indicates the model yields few false positives, but the lower sensitivity (70%) indicates that not all virulence determinants may have been captured as a result of the small size of the data set. Noise and experimental errors in the data are also expected to reduce the classification performance. The performance of the model compares well with other models relating viral mutations to phenotype, e.g., sensitivity in the range 40 to 100% (70% was typical) and a specificity of ca. 70% have been reported for HIV drug resistance models (42).

In conclusion, we have collated experimental data about the virulence of HPAI H5N1 strains in mammals and used it to identify its genetic basis by inferring a BGM of whole-genome mutations affecting the phenotype. We used data from strains isolated over a period of at least 10 years and analyzed the results obtained in 27 studies so our conclusions are general rather than specific to a single experimental system. The resulting statistical model reveals the polygenic nature of virulence in mammals: mutations at PB1-317, PB2-355, NS1-92, NS1-228, NS1-195 and HA-102(86) are directly associated with the trait. These sites split into two semi-independent groups, possibly indicating separate virulence mechanisms: PB1-317, PB2-355, NS1-92, and NS1-228 mutations were strongly correlated with each other, as were NS1-195 and HA-102(86). We also found strong correlation patterns in the PB1, PB2, and PA sites; the antigenic sites and a variable glycosylation site in HA; and the HA and NS sites. The results highlight the importance of the polymerase and nonstructural internal genes in addition to HA in determining the virulence of HPAI H5N1 in mammals and identify potential targets for intervention in mammalian H5N1 infections.

ACKNOWLEDGMENTS

This study was funded by the Biotechnology and Biological Sciences Research Council (Bb/B009670/1). In addition, this research was supported in part by the National Institutes of Health (AI43638, AI47745, and AI57167), the University of California Universitywide AIDS Research Program (grant number IS02-SD-701), and a University of California, San Diego, Center for AIDS Research/NIAID Developmental Award (AI36214).

REFERENCES

1. **Alexander, D. J.** 2000. A review of avian influenza in different bird species. Vet. Microbiol. **74:**3–13.
2. **Amonsin, A., T. Songserm, S. Chutinimitkul, R. Jam-on, N. Sae-Heng, N. Pariyothorn, S. Payungporn, A. Theamboonlers, and Y. Poovorawan.** 2007. Genetic analysis of influenza A virus (H5N1) derived from domestic cat and dog in Thailand. Arch. Virol. **152:**1925–1933.

3. **Baigent, S. J., and J. W. McCauley.** 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.
4. **Bao, Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman.** 2008. The influenza virus resource at the National Center for Biotechnology Information. J. Virol. **82:**596–601.
5. **Bender, C., H. Hall, J. Huang, A. Klimov, N. Cox, A. Hay, V. Gregory, K. Cameron, W. Lim, and K. Subbarao.** 1999. Characterization of the surface proteins of influenza A (H5N1) viruses isolated from humans in 1997–1998. Virology **254:**115–123.
6. **Bornholdt, Z. A., and B. V. V. Prasad.** 2006. X-ray structure of influenza virus NS1 effector domain. Nat. Struct. Mol. Biol. **13:**559–560.
7. **Bornholdt, Z. A., and B. V. V. Prasad.** 2008. X-ray structure of NS1 from a highly pathogenic H5N1 influenza virus. Nature **456:**985–988.
8. **Bright, R. A., T. M. Ross, K. Subbarao, H. L. Robinson, and J. M. Katz.** 2003. Impact of glycosylation on the immunogenicity of a DNA-based influenza H5 HA vaccine. Virology **308:**270–278.
9. **Cauthen, A. N., D. E. Swayne, S. Schultz-Cherry, M. L. Perdue, and D. L. Suarez.** 2000. Continued circulation in china of highly pathogenic avian influenza viruses encoding the hemagglutinin gene associated with the 1997 H5N1 outbreak in poultry and humans. J. Virol. **74:**6592–6599.
10. **Chen, H., R. A. Bright, K. Subbarao, C. Smith, N. J. Cox, J. M. Katz, and Y. Matsuoka.** 2007. Polygenic virulence factors involved in pathogenesis of 1997 Hong Kong H5N1 influenza viruses in mice. Virus Res. **128:**159–163.
11. **Chen, H., G. Deng, Z. Li, G. Tian, Y. Li, P. Jiao, L. Zhang, Z. Liu, R. G. Webster, and K. Yu.** 2004. The evolution of H5N1 influenza viruses in ducks in southern China. Proc. Natl. Acad. Sci. USA **101:**10452–10457.
12. **Chen, H., G. J. D. Smith, K. S. Li, J. Wang, X. H. Fan, J. M. Rayner, D. Vijaykrishna, J. X. Zhang, L. J. Zhang, C. T. Guo, C. L. Cheung, K. M. Xu, L. Duan, K. Huang, K. Qin, Y. H. C. Leung, W. L. Wu, H. R. Lu, Y. Chen, N. S. Xia, T. S. P. Naipospos, K. Y. Yuen, S. S. Hassan, S. Bahri, T. D. Nguyen, R. G. Webster, J. S. M. Peiris, and Y. Guan.** 2006. Establishment of multiple sublineages of H5N1 influenza virus in Asia: implications for pandemic control. Proc. Natl. Acad. Sci. USA **103:**2845–2850.
13. **Chen, H., G. J. D. Smith, S. Y. Zhang, K. Qin, J. Wang, K. S. Li, R. G. Webster, J. S. M. Peiris, and Y. Guan.** 2005. Avian flu H5N1 virus outbreak in migratory waterfowl. Nature **436:**191–192.
14. **Chen, R., and E. C. Holmes.** 2006. Avian influenza virus exhibits rapid evolutionary dynamics. Mol. Biol. Evol. **23:**2336–2341.
15. **Claas, E. C. J., A. D. M. E. Osterhaus, R. van Beek, J. C. De Jong, G. F. Rimmelzwaan, D. A. Senne, S. Krauss, K. F. Shortridge, and R. G. Webster.** 1998. Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. Lancet **351:**472–477.
16. **Cooper, G. F., and E. Herskovits.** 1992. A Bayesian method for the induction of probabilistic networks from data. Machine Learning **9:**309–347.
17. **de Jong, M. D., C. P. Simmons, T. T. Thanh, V. M. Hien, G. J. D. Smith, T. N. B. Chau, D. M. Hoang, N. Van Vinh Chau, T. H. Khanh, V. C. Dong, P. T. Qui, B. Van Cam, D. Q. Ha, Y. Guan, J. S. M. Peiris, N. T. Chinh, T. T. Hien, and J. Farrar.** 2006. Fatal outcome of human influenza A (H5N1) is associated with high viral load and hypercytokinemia. Nat. Med. **12:**1203–1207.
18. **Friedman, N.** 2004. Inferring cellular networks using probabilistic graphical models. Science **303:**799–805.
19. **Friedman, N., and D. Koller.** 2003. Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks. Machine Learning **50:**95–125.
20. **Govorkova, E. A., J. E. Rehg, S. Krauss, H.-L. Yen, Y. Guan, M. Peiris, T. D. Nguyen, T. H. Hanh, P. Puthavathana, H. T. Long, C. Buranathai, W. Lim, R. G. Webster, and E. Hoffmann.** 2005. Lethality to ferrets of H5N1 influenza viruses isolated from humans and poultry in 2004. J. Virol. **79:**2191–2198.
21. **Guan, Y., J. S. M. Peiris, A. S. Lipatov, T. M. Ellis, K. C. Dyrting, S. Krauss, L. J. Zhang, R. G. Webster, and K. F. Shortridge.** 2002. Emergence of multiple genotypes of H5N1 avian influenza viruses in Hong Kong SAR. Proc. Natl. Acad. Sci. USA **99:**8950–8955.
22. **Guan, Y., L. L. M. Poon, C. Y. Cheung, T. M. Ellis, W. Lim, A. S. Lipatov, K. H. Chan, K. M. Sturm-Ramirez, C. L. Cheung, Y. H. C. Leung, K. Y. Yuen, R. G. Webster, and J. S. M. Peiris.** 2004. H5N1 influenza: a protean pandemic threat. Proc. Natl. Acad. Sci. USA **101:**8156–8161.
23. **Ha, Y., D. J. Stevens, J. J. Skehel, and D. C. Wiley.** 2001. X-ray structures of H5 avian and H9 swine influenza virus hemagglutinins bound to avian and human receptor analogs. Proc. Natl. Acad. Sci. USA **98:**11181–11186.
24. **Hatta, M., P. Gao, P. Halfmann, and Y. Kawaoka.** 2001. Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. Science **293:**1840–1842.
25. **Hatta, M., Y. Hatta, J. H. Kim, S. Watanabe, K. Shinya, T. Nguyen, P. S. Lien, Q. M. Le, and Y. Kawaoka.** 2007. Growth of H5N1 influenza a viruses in the upper respiratory tracts of mice. PLoS Pathog. **3:**e133.
26. **Heikkinen, L. S., A. Kazlauskas, K. Melen, R. Wagner, T. Ziegler, I. Julkunen, and K. Saksela.** 2008. Avian and 1918 Spanish influenza A virus NS1 proteins bind to Crk/CrkL SH3 domains to activate host cell signalling. J. Biol. Chem. **283:**5719–5727.
27. **Horimoto, T., and Y. Kawaoka.** 1994. Reverse genetics provides direct evidence for a correlation of hemagglutinin cleavability and virulence of an avian influenza A virus. J. Virol. **68:**3120–3128.
28. **Hulse-Post, D. J., J. Franks, K. Boyd, R. Salomon, E. Hoffmann, H. L. Yen, R. J. Webby, D. Walker, T. D. Nguyen, and R. G. Webster.** 2007. Molecular changes in the polymerase genes (PA and PB1) associated with high pathogenicity of H5N1 influenza virus in mallard ducks. J. Virol. **81:**8515–8524.
29. **Hulse-Post, D. J., K. M. Sturm-Ramirez, J. Humberd, P. Seiler, E. A. Govorkova, S. Krauss, C. Scholtissek, P. Puthavathana, C. Buranathai, T. D. Nguyen, H. T. Long, T. S. P. Naipospos, H. Chen, T. M. Ellis, Y. Guan, J. S. M. Peiris, and R. G. Webster.** 2005. Role of domestic ducks in the propagation and biological evolution of highly pathogenic H5N1 influenza viruses in Asia. Proc. Natl. Acad. Sci. USA **102:**10682–10687.
30. **Ito, T., Y. Suzuki, A. Takada, A. Kawamoto, K. Otsuki, H. Masuda, M. Yamada, T. Suzuki, H. Kida, and Y. Kawaoka.** 1997. Differences in sialic acid-galactose linkages in the chicken egg amnion and allantois influence human influenza virus receptor specificity and variant selection. J. Virol. **71:**3357–3362.
31. **Jackson, D., M. J. Hossain, D. Hickman, D. R. Perez, and R. A. Lamb.** 2008. A new influenza virus virulence determinant: the NS1 protein four C-terminal residues modulate pathogenicity. Proc. Natl. Acad. Sci. USA **105:**4381–4386.
32. **Jansen, R., N. Lan, J. Qian, and M. Gerstein.** 2002. Integration of genomic datasets to predict protein complexes in yeast. J. Struct. Funct. Genomics **2:**71–81.
33. **Jiao, P., G. Tian, Y. Li, G. Deng, Y. Jiang, C. Liu, W. Liu, Z. Bu, Y. Kawaoka, and H. Chen.** 2008. A single-amino-acid substitution in the NS1 protein changes the pathogenicity of H5N1 avian influenza viruses in mice. J. Virol. **82:**1146–1154.
34. **Katz, J. M., X. Lu, T. M. Tumpey, C. B. Smith, M. W. Shaw, and K. Subbarao.** 2000. Molecular correlates of influenza A H5N1 virus pathogenesis in mice. J. Virol. **74:**10807–10810.
35. **Keawcharoen, J., K. Oraveerakul, T. Kuiken, R. A. Fouchier, A. Amonsin, S. Payungporn, S. Noppornpanth, S. Wattanodorn, A. Theambooniers, R. Tantilertcharoen, R. Pattanarangsan, N. Arya, P. Ratanakorn, D. M. Osterhaus, and Y. Poovorawan.** 2004. Avian influenza H5N1 in tigers and leopards. Emerg. Infect. Dis. **10:**2189–2191.
36. **Kuiken, T., G. Rimmelzwaan, D. van Riel, G. van Amerongen, M. Baars, R. Fouchier, and A. Osterhaus.** 2004. Avian H5N1 influenza in cats. Science **306:**241.
37. **Lauritzen, S. L., and N. A. Sheehan.** 2003. Graphical models for genetic analyses. Stat. Sci. **18:**489–514.
38. **Le, Q. M., Y. Sakai-Tagawa, M. Ozawa, M. Ito, and Y. Kawaoka.** 2009. Selection of H5N1 influenza virus PB2 during replication in humans. J. Virol. **83:**5278–5281.
39. **Lee, C.-W., D. L. Suarez, T. M. Tumpey, H.-W. Sung, Y.-K. Kwon, Y.-J. Lee, J.-G. Choi, S.-J. Joh, M.-C. Kim, E.-K. Lee, J.-M. Park, X. Lu, J. M. Katz, E. Spackman, D. E. Swayne, and J.-H. Kim.** 2005. Characterization of highly pathogenic H5N1 avian influenza A viruses isolated from South Korea. J. Virol. **79:**3692–3702.
40. **Lee, I., S. V. Date, A. T. Adai, and E. M. Marcotte.** 2004. A probabilistic functional network of yeast genes. Science **306:**1555–1558.
41. **Lee, M. S., M. C. Deng, Y. J. Lin, C. Y. Chang, H. K. Shieh, J. Z. Shiau, and C. C. Huang.** 2007. Characterization of an H5N1 avian influenza virus from Taiwan. Vet. Microbiol. **124:**193–201.
42. **Leigh Brown, A. J., S. D. W. Frost, B. Good, E. S. Daar, V. Simon, M. Markowitz, A. C. Collier, E. Connick, B. Conway, J. B. Margolick, J.-P. Routy, J. Corbeil, N. S. Hellmann, D. D. Richman, and S. J. Little.** 2004. Genetic basis of hypersusceptibility to protease inhibitors and low replicative capacity of human immunodeficiency virus type 1 strains in primary infection. J. Virol. **78:**2242–2246.
43. **Li, K. S., Y. Guan, J. Wang, G. J. D. Smith, K. M. Xu, L. Duan, A. P. Rahardjo, P. Puthavathana, C. Buranathai, T. D. Nguyen, A. T. S. Estoepangestie, A. Chaisingh, P. Auewarakul, H. T. Long, N. T. H. Hanh, R. J. Webby, L. L. M. Poon, H. Chen, K. F. Shortridge, K. Y. Yuen, R. G. Webster, and J. S. M. Peiris.** 2004. Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. Nature **430:**209–213.
44. **Li, M., and B. Wang.** 2007. Homology modeling and examination of the effect of the D92E mutation on the H5N1 nonstructural protein NS1 effector domain. J. Mol. Model. **13:**1237–1244.
45. **Li, Z., Y. Jiang, P. Jiao, A. Wang, F. Zhao, G. Tian, X. Wang, K. Yu, Z. Bu, and H. Chen.** 2006. The NS1 gene contributes to the virulence of H5N1 avian influenza viruses. J. Virol. **80:**11115–11123.
46. **Lipatov, A. S., S. Andreansky, R. J. Webby, D. J. Hulse, J. E. Rehg, S. Krauss, D. R. Perez, P. C. Doherty, R. G. Webster, and M. Y. Sangster.** 2005. Pathogenesis of Hong Kong H5N1 influenza virus NS gene reassortants in mice: the role of cytokines and B- and T-cell responses. J. Gen. Virol. **86:**1121–1130.
47. **Lipatov, A. S., S. Krauss, Y. Guan, M. Peiris, J. E. Rehg, D. R. Perez, and R. G. Webster.** 2003. Neurovirulence in mice of H5N1 influenza virus genotypes isolated from Hong Kong poultry in 2001. J. Virol. **77:**3816–3823.
48. **Long, J.-X., D.-X. Peng, Y.-L. Liu, Y.-T. Wu, and X.-F. Liu.** 2008. Virulence

of H5N1 avian influenza virus enhanced by a 15-nucleotide deletion in the viral nonstructural gene. Virus Genes **36:**471–478.

49. **Ludwig, S., U. Schultz, J. Mandler, W. M. Fitch, and C. Scholtissek.** 1991. Phylogenetic relationship of the nonstructural (NS) genes of influenza A viruses. Virology **183:**566–577.
50. **Maines, T. R., X. H. Lu, S. M. Erb, L. Edwards, J. Guarner, P. W. Greer, D. C. Nguyen, K. J. Szretter, L.-M. Chen, P. Thawatsupha, M. Chittaganpitch, S. Waicharoen, D. T. Nguyen, T. Nguyen, H. H. T. Nguyen, J.-H. Kim, L. T. Hoang, C. Kang, L. S. Phuong, W. Lim, S. Zaki, R. O. Donis, N. J. Cox, J. M. Katz, and T. M. Tumpey.** 2005. Avian influenza (H5N1) viruses isolated from humans in Asia in 2004 exhibit increased virulence in mammals. J. Virol. **79:**11788–11800.
51. **Marcotte, E. M., I. Xenarios, and D. Eisenberg.** 2001. Mining literature for protein-protein interactions. Bioinformatics **17:**359–363.
52. **Matrosovich, M., N. Zhou, Y. Kawaoka, and R. Webster.** 1999. The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distinguishable properties. J. Virol. **73:**1146–1155.
53. **Min, J.-Y., S. Li, G. C. Sen, and R. M. Krug.** 2007. A site on the influenza A virus NS1 protein mediates both inhibition of PKR activation and temporal regulation of viral RNA synthesis. Virology **363:**236–243.
54. **Mitnaul, L. J., M. N. Matrosovich, M. R. Castrucci, A. B. Tuzikov, N. V. Bovin, D. Kobasa, and Y. Kawaoka.** 2000. Balanced hemagglutinin and neuraminidase activities are critical for efficient replication of influenza A virus. J. Virol. **74:**6015–6020.
55. **Mushtaq, M., H. Juan, P. Jiang, Y. Li, T. Li, Y. Du, and M. Mukhtar.** 2008. Complete genome analysis of a highly pathogenic H5N1 influenza A virus isolated from a tiger in China. Arch. Virol. **153:**1569–1574.
56. **Myers, C., D. Barrett, M. Hibbs, C. Huttenhower, and O. Troyanskaya.** 2006. Finding function: evaluation methods for functional genomic data. BMC Genomics **7:**187.
57. **Obenauer, J. C., J. Denson, P. K. Mehta, X. Su, S. Mukatira, D. B. Finkelstein, X. Xu, J. Wang, J. Ma, Y. Fan, K. M. Rakestraw, R. G. Webster, E. Hoffmann, S. Krauss, J. Zheng, Z. Zhang, and C. W. Naeve.** 2006. Large-scale sequence analysis of avian influenza isolates. Science **311:**1576–1580.
58. **Olsen, B., V. J. Munster, A. Wallensten, J. Waldenstrom, A. D. M. E. Osterhaus, and R. A. M. Fouchier.** 2006. Global patterns of influenza A virus in wild birds. Science **312:**384–388.
59. **Pearl, J.** 1986. Fusion, propagation, and structuring in belief networks. Artif. Intell. **29:**241–288.
60. **Poon, A. F., F. I. Lewis, S. D. Frost, and S. L. Kosakovsky Pond.** 2008. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. Bioinformatics **24:**1949–1950.
61. **Poon, A. F., F. I. Lewis, S. L. Pond, and S. D. Frost.** 2007. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. PLoS Comput. Biol. **3:**e231.
62. **Rott, R.** 1980. Genetic determinants for infectivity and pathogenicity of influenza viruses. Phil. Trans. R. Soc. London B Biol. Sci. **288:**393–399.
63. **Rott, R.** 1992. The pathogenic determinant of influenza virus. Vet. Microbiol. **33:**303–310.
64. **Salomon, R., J. Franks, E. A. Govorkova, N. A. Ilyushina, H.-L. Yen, D. J. Hulse-Post, J. Humberd, M. Trichet, J. E. Rehg, R. J. Webby, R. G. Webster, and E. Hoffmann.** 2006. The polymerase complex genes contribute to the high virulence of the human H5N1 influenza virus isolate A/Vietnam/1203/04. J. Exp. Med. **203:**689–697.
65. **Searls, D. B.** 2005. Data integration: challenges for drug discovery. Nat. Rev. Drug Discov. **4:**45–58.
66. **Sebastiani, P., M. F. Ramoni, V. Nolan, C. T. Baldwin, and M. H. Steinberg.** 2005. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nat. Genet. **37:**435–440.
67. **Seo, S. H., E. Hoffmann, and R. G. Webster.** 2002. Lethal H5N1 influenza viruses escape host antiviral cytokine responses. Nat. Med. **8:**950–954.
68. **Smith, G. J. D., T. S. P. Naipospos, T. D. Nguyen, M. D. de Jong, D. Vijaykrishna, T. B. Usman, S. S. Hassan, T. V. Nguyen, T. V. Dao, N. A. Bui, Y. H. C. Leung, C. L. Cheung, J. M. Rayner, J. X. Zhang, L. J. Zhang, L. L. M. Poon, K. S. Li, V. C. Nguyen, T. T. Hien, J. Farrar, R. G. Webster, H. Chen, J. S. M. Peiris, and Y. Guan.** 2006. Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. Virology **350:**258–268.
69. **Stark, C., B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers.** 2006. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. **34:**D535–D539.
70. **Steel, J., A. C. Lowen, S. Mubareka, and P. Palese.** 2009. Transmission of influenza virus in a mammalian host is increased by PB2 amino acids 627K or 627E/701N. PLoS Pathog. **5:**e1000252.
71. **Sturm-Ramirez, K. M., T. Ellis, B. Bousfield, L. Bissett, K. Dyrting, J. E. Rehg, L. Poon, Y. Guan, M. Peiris, and R. G. Webster.** 2004. Reemerging H5N1 influenza viruses in Hong Kong in 2002 are highly pathogenic to ducks. J. Virol. **78:**4892–4901.
72. **Suarez, D. L., M. L. Perdue, N. Cox, T. Rowe, C. Bender, J. Huang, and D. E. Swayne.** 1998. Comparisons of highly virulent H5N1 influenza A viruses isolated from humans and chickens from Hong Kong. J. Virol. **72:**6678–6688.
73. **Subbarao, E. K., W. London, and B. R. Murphy.** 1993. A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. J. Virol. **67:**1761–1764.
74. **Subbarao, K., A. Klimov, J. Katz, H. Regnery, W. Lim, H. Hall, M. Perdue, D. Swayne, C. Bender, J. Huang, M. Hemphill, T. Rowe, M. Shaw, X. Xu, K. Fukuda, and N. Cox.** 1998. Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness. Science **279:**393–396.
75. **Subbarao, K., and M. W. Shaw.** 2000. Molecular aspects of avian influenza (H5N1) viruses isolated from humans. Rev. Med. Virol. **10:**337–348.
76. **Thiry, E., A. Zicola, D. Addie, H. Egberink, K. Hartmann, H. Lutz, H. Poulet, and M. C. Horzinek.** 2007. Highly pathogenic avian influenza H5N1 virus in cats and other carnivores. Vet. Microbiol. **122:**25–31.
77. **Vijaykrishna, D., J. Bahl, S. Riley, L. Duan, J. X. Zhang, H. Chen, J. S. M. Peiris, G. J. D. Smith, and Y. Guan.** 2008. Evolutionary dynamics and emergence of panzootic H5N1 influenza viruses. PLoS Pathog. **4:**e1000161.
78. **Wagner, R., M. Matrosovich, and H. D. Klenk.** 2002. Functional balance between haemagglutinin and neuraminidase in influenza virus infections. Rev. Med. Virol. **12:**159–166.
79. **Wagner, R., T. Wolff, A. Herwig, S. Pleschka, and H. D. Klenk.** 2000. Interdependence of hemagglutinin glycosylation and neuraminidase as regulators of influenza virus growth: a study by reverse genetics. J. Virol. **74:**6316–6323.
80. **Wan, X.-F., T. Nguyen, C. T. Davis, C. B. Smith, Z.-M. Zhao, M. Carrel, K. Inui, H. T. Do, D. T. Mai, S. Jadhao, A. Balish, B. Shu, F. Luo, M. Emch, Y. Matsuoka, S. E. Lindstrom, N. J. Cox, C. V. Nguyen, A. Klimov, and R. O. Donis.** 2008. Evolution of highly pathogenic H5N1 avian influenza viruses in Vietnam between 2001 and 2007. PLoS ONE **3:**e3462.
81. **Webster, R. G., S. Krauss, D. Hulse-Post, and K. Sturm-Ramirez.** 2007. Evolution of influenza A viruses in wild birds. J. Wildl. Dis. **43:**S1–S6.
82. **WHO/OIE/FAO H5N1 Evolution Working Group.** 2008. Toward a unified nomenclature system for highly pathogenic avian influenza virus (H5N1). Emerg. Infect. Dis. **14:**e1.
83. **Wu, W. L., Y. Chen, P. Wang, W. Song, S.-Y. Lau, J. M. Rayner, G. J. D. Smith, R. G. Webster, J. S. M. Peiris, T. Lin, N. Xia, Y. Guan, and H. Chen.** 2008. Antigenic profile of avian H5N1 viruses in Asia from 2002 to 2007. J. Virol. **82:**1798–1807.
84. **Xu, X., K. Subbarao, N. J. Cox, and Y. Guo.** 1999. Genetic characterization of the pathogenic influenza A/Goose/Guangdong/1/96 (H5N1) virus: similarity of its hemagglutinin gene to those of H5N1 viruses from the 1997 outbreaks in Hong Kong. Virology **261:**15–19.
85. **Yen, H. L., A. S. Lipatov, N. A. Ilyushina, E. A. Govorkova, J. Franks, N. Yilmaz, A. Douglas, A. Hay, S. Krauss, J. E. Rehg, E. Hoffmann, and R. G. Webster.** 2007. Inefficient transmission of H5N1 influenza viruses in a ferret contact model. J. Virol. **81:**6890–6898.
86. **Zitzow, L. A., T. Rowe, T. Morken, W. J. Shieh, S. Zaki, and J. M. Katz.** 2002. Pathogenesis of avian influenza A (H5N1) viruses in ferrets. J. Virol. **76:** 4420–4429.

DOI: 10.1534/genetics.109.106021

Phylodynamics of Infectious Disease Epidemics

Erik M. Volz,*,†,[1] Sergei L. Kosakovsky Pond,‡ Melissa J. Ward,§ Andrew J. Leigh Brown§ and Simon D. W. Frost**

**Department of Epidemiology, University of Michigan, Ann Arbor, Michigan 48109, †Department of Pathology and ‡Department of Medicine, University of California, La Jolla, California 92093, §School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JR, United Kingdom and **Department of Veterinary Medicine, University of Cambridge, Cambridge CB3 0ES, United Kingdom*

Manuscript received June 8, 2009
Accepted for publication September 11, 2009

ABSTRACT

We present a formalism for unifying the inference of population size from genetic sequences and mathematical models of infectious disease in populations. Virus phylogenies have been used in many recent studies to infer properties of epidemics. These approaches rely on coalescent models that may not be appropriate for infectious diseases. We account for phylogenetic patterns of viruses in susceptible–infected (SI), susceptible–infected–susceptible (SIS), and susceptible–infected–recovered (SIR) models of infectious disease, and our approach may be a viable alternative to demographic models used to reconstruct epidemic dynamics. The method allows epidemiological parameters, such as the reproductive number, to be estimated directly from viral sequence data. We also describe patterns of phylogenetic clustering that are often construed as arising from a short chain of transmissions. Our model reproduces the moments of the distribution of phylogenetic cluster sizes and may therefore serve as a null hypothesis for cluster sizes under simple epidemiological models. We examine a small cross-sectional sample of human immunodeficiency (HIV)-1 sequences collected in the United States and compare our results to standard estimates of effective population size. Estimated prevalence is consistent with estimates of effective population size and the known history of the HIV epidemic. While our model accurately estimates prevalence during exponential growth, we find that periods of decline are harder to identify.

COALESCENT theory has found wide applications for inference of viral phylogenies (Nee *et al.* 1996; Rosenberg and Nordborg 2002; Drummond *et al.* 2005) and estimation of epidemic prevalence (Yusim *et al.* 2001; Robbins *et al.* 2003; Wilson *et al.* 2005), yet there have been few attempts to formally integrate coalescent theory with standard epidemiological models (Pybus *et al.* 2001; Goodreau 2006). While epidemiological models such as susceptible–infected–recovered (SIR) consider the dynamics of an entire population going forward in time, the coalescent theory operates on a small sample of an infected subpopulation and models the merging of lineages backward in time until a common ancestor has been reached. The original coalescent theory was based on a population of constant size with discrete generations (Kingman 1982a,b). Numerous extensions have been made for populations with overlapping generations in continuous time, exponential or logistic growth (Griffiths and Tavare 1994), and stochastically varying size (Kaj and Krone 2003). However, infectious disease epidemics are a special case of a variable size population, often characterized by early explosive growth followed by decline that leads to extinction or an endemic steady state.

If superinfection is rare and if mutation is fast relative to epidemic growth, each lineage in a phylogenetic tree corresponds to a single infected individual with its own unique viral population. An infection event viewed in reverse time is equivalent to the coalescence of two lineages and every transmission of the virus between hosts can generate a new branch in the phylogeny of consensus viral isolates from infected individuals. Recently diverged sequences should represent transmissions in the recent past, and branches close to the root of a tree should represent transmissions from long ago. Consequently, branching patterns provide information about the frequency of transmissions over time (Wilson *et al.* 2005). The correspondence between transmission and phylogenetic branching is easiest to detect for viruses such as human immunodeficiency virus (HIV) and hepatitis C virus that have a high mutation rate relative to dispersal. Underlying SIR dynamics also apply to other pathogens, although in some cases it may be more difficult to reconstruct the transmission history.

We examined the properties of viral phylogenies generated by the most common epidemiological models based on ordinary differential equations (ODEs).

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.109.106021/DC1.

[1] *Corresponding author:* Department of Epidemiology, University of Michigan, M5073, SPH II, 1415 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: erik@erikvolz.info

We are able to fit epidemiological models to a reconstructed phylogeny for sampled viral sequence data and make inferences regarding the size of the corresponding infected population. Our solution takes the form of an ODE analogous to those used to track epidemic prevalence and thereby provides a convenient link between commonly used epidemiological models and phylodynamics. Virtually all coalescent theory to date has been expressed in terms of integer-valued stochastic processes. Our motivation for using differential equations to describe the coalescent process is a desire to formalize a link with standard epidemiological models that are also expressed in terms of differential equations.

We use our method to calculate the distribution of coalescent times for samples of viral sequences, fit SIR models to a viral phylogeny, and calculate median time to the most recent common ancestor (MRCA) of the sample. Our method also provides equations that describe the time evolution of the cluster size distribution (CSD)—the distribution of the number of descendants of a lineage over time. Clusters of related virus are often interpreted as epidemiologically linked. For example, clusters of acute HIV infections may represent short transmission chains between high-risk individuals (Yerly *et al.* 2001; Hue *et al.* 2005; Pao *et al.* 2005; Goodreau 2006; Brenner *et al.* 2007; Drumright and Frost 2008; Lewis *et al.* 2008). Because our model reproduces the moments of the cluster size distribution, it can be used to predict the level of clustering as a function of epidemiological conditions. The moments could be directly compared to empirical values or they could be used to reconstruct the entire CSD, whereupon standard statistical tests could be used for comparing distributions.

Although our equations describe the macroscopic properties of the population distribution of cluster sizes, we generalize our method to the case of a small cross-sectional sample of sequences. This allows us to develop a likelihood-based approach to fitting SIR models to observed sequences.

By considering variable degrees of incidence and the size of the infected population, our solution sheds light on the relationship between coalescent rates and epidemic dynamics. Coalescent rates are low near peak prevalence, but higher when there is a large ratio of incidence to prevalence. This can occur early on, when the epidemic is entering its expansion phase, as well as late if the epidemic has multiple periods of growth.

METHODS

Consider a population of size N comprising susceptible ($\mathcal{S}$), infected ($\mathcal{I}$), and recovered ($\mathcal{R}$) individuals. The deterministic limiting behavior of $S=|\mathcal{S}|/N$, $I=|\mathcal{I}|/N$, and $R=|\mathcal{R}|/N$ as $N\rightarrow\infty$ and with all variables $\gg 1/N$ is described by a set of coupled ordinary differential equations, with time-dependent rates of change from state X to state Y denoted as $f_{XY}(t)$. For instance, the classical mass-action SIR model

$$\dot{S}=-\beta SI,\ \dot{I}=\beta SI-\gamma I,\ \dot{R}=\gamma I \qquad (1)$$

(Kermack and McKendrick 1927; Bailey 1975; Anderson and May 1991) is obtained by setting $f_{SI}(t)=\beta S(t)I(t)$, $f_{IR}(t)=\gamma I(t)$, and all other rates to 0. We omit the explicit dependence of terms on time when it is unambiguous.

Classical coalescent inference operates on a small subsample of the larger evolving population, taken at a single time point, and infers properties of the population at an earlier time point; *e.g.*, What is the expected number of lineages at a given time *t*? Here, we denote the time of sampling by T and consider the evolution of the population backward in time toward time $t=0$. While this differs from the conventional temporal notation for coalescent theory (where the sampling, or present, time is denoted 0, and as we move backward t denotes the number of years before the present), it allows us to develop a system of equations that link coalescent inference with standard epidemiological models.

We apply the coalescent model to the population of infecteds ($\mathcal{I}$) and draw upon the dynamical system to provide parameters such as the rate of lineage coalescence. The practical questions that we seek to address include the following:

If n individuals are sampled at time T, how many lineages exist at time $t\leq T$?

How many lineages extant at time t have surviving progeny at time T? We define *progeny* of a viral lineage extant from time $t\leq T$ as those individuals infected at time T whose virus can be traced back to that viral lineage at time t. For instance, in Figure 1, from $t=t_1$ the progeny of lineage 6 has four individuals (5, 6, 8, and 9), but from $t=t_2$ the progeny of lineage 6 consists of only 5 and 6.

Can we describe the distribution of the number of progeny from time t (a time t cluster), $\mathbf{X}(t)$, using its distributional moments? For instance, in Figure 1, at time $t=t_2$ this distribution is given by (2, 2, 2), while for $t=t_1$ the distribution is (4, 2).

Note that a transmission does not always result in an observable coalescent event depending on which lineages expire due to recovery or are not sampled (*e.g.*, the transmission from 7 to 10 in Figure 1), and a transmission to an individual that recovers may still correspond to a coalescent event if that person transmits prior to recovering (*e.g.*, the transmission from 6 to 7 in Figure 1).

Coalescent model for SIR epidemics: In an SIR epidemic, a branch in the tree corresponds to a transmission event, and as a lineage is traced backward in time, it traverses multiple infected hosts. A recovery event before the sample time T does not alter the number of lineages with progeny because no progeny

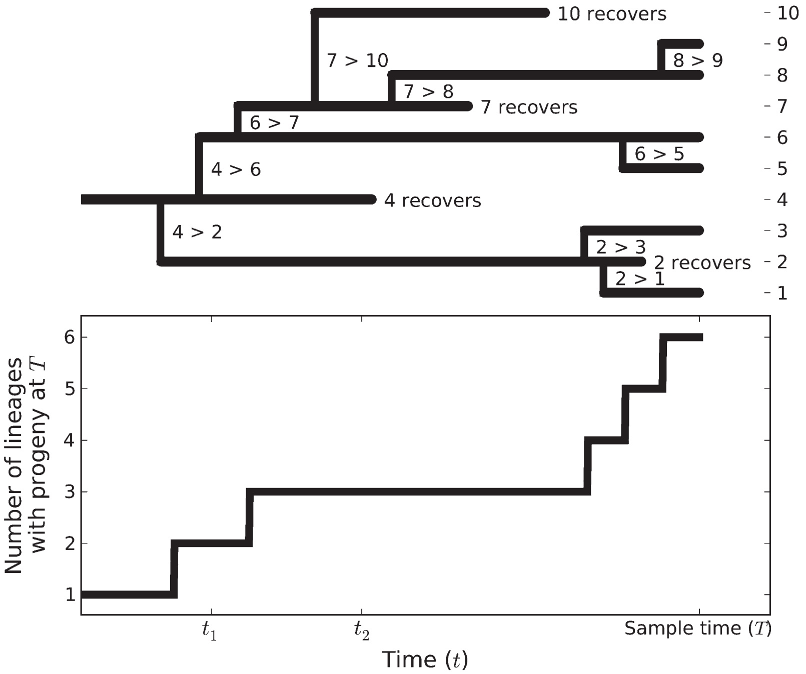


FIGURE 1.—An example of a phylogeny that could be generated by an epidemic process. The number of lineages at time t for a population observed at time T is plotted below. A branch in the tree corresponds to a transmission event, and as a lineage is traced backward in time, it traverses multiple infected hosts.

of this individual can be sampled at a later time. In a standard coalescent model, n lineages merge in reverse time at a rate proportional to $\binom{n}{2}$. Given that a coalescent event occurs among the individuals in $\mathcal{I}$, the probability of observing it among the n observed lineages is

$$\binom{n}{2}\Big/\binom{|\mathcal{I}|}{2} = \frac{n(n-1)}{|\mathcal{I}|\,(|\mathcal{I}|-1)}.$$

We introduce the dimensionless variable $A(t;\ T)$, which is the fraction of the population at t with sampled progeny extant at T. $A(t;\ T)$ is proportional to the number of ancestors of a sample of sequences and is analogous to the integer-valued ancestor function used in standard coalescent theory (GRIFFITHS and TAVARE 1994). We consider how A evolves as t moves into the past, with T fixed.

If a fraction ϕ of the infected population is sampled at time T, then we observe a number $n = \phi|\mathcal{I}(T)|$ lineages. Initially, $t = T$, and $A(T;\ T) = \phi I$ (the ancestor of each sequence is itself). The sample fraction ϕ is not always known, but if $\phi = 1$, our solution will describe the evolution of the fraction of extant lineages for the entire population.

Using the definition of A and assuming $A \gg 1/N$, the probability of a transmission event causing a coalescent event to be observed in our sample is

$$p_c(t;\ T) = \lim_{N\to\infty}\frac{\binom{A(t;\ T)N}{2}}{\binom{NI(t)}{2}} = \left(\frac{A(t;\ T)}{I(t)}\right)^2.$$

The rate of coalescence for a sample of sequences is analogous to the rate of change of the ancestor function, A. We can write the coalescence rate for the sample of sequences as the product of the number of transmissions per unit time, $f_{SI}(t)$ and the probability p_c that a transmission results in a coalescence being observed in our sample. The ancestor function $A(t;\ T)$ can be found by integrating the following backward ordinary differential equation from time T:

$$-\frac{dA}{dt} := \dot{A} = -f_{SI}p_c = -f_{SI}\left(\frac{A}{I}\right)^2. \qquad (2)$$

This equation works even when $\phi = 1$, in which case A represents the number of ancestors of the entire population of infecteds observed at time T.

Cluster size distribution: Let $\mathbf{X}_1(t;\ T)$ denote the number of progeny at T of a random infected host from time $t \leq T$, given that such progeny exist. We denote the expected value of $\mathbf{X}_1$ by $x_1(t;\ T)$ and interpret it as the *mean cluster size* from time t. $\mathbf{X}_2(t;\ T)$ [and $x_2 = E(\mathbf{X}_2)$] is a random variable that describes the size of the cluster if it is selected with probability proportional to the cluster's size. This is the same distribution of cluster sizes as if we select an infected at time T and determine the size of the cluster to which that infected belongs.

Below, we show that x_1 and x_2 can be found by integrating the ordinary differential equations

$$\dot{x}_1(t;\ T) = f_{SI}(t)I(T)/I(t)^2, \qquad (3)$$

$$\dot{x}_2 = 2\dot{x}_1 \qquad (4)$$

backward in time from T with initial prevalence $I(T)$ taken from the epidemic model. Also, initially (at $t = T$), all cluster sizes are unity, and $x_1(T;\ T) = x_2(T;\ T) = 1$.

The set of infecteds $\mathcal{I}(T)$ is distributed across a number $A(t;\ T)N$ clusters, and for any $0 \leq t \leq T$, the average number of infecteds per time-t cluster is $I(T)/A(t;\ T)$. This implies

$$A(t;\ T) = I(T)/x_1(t;\ T). \tag{5}$$

Evaluating the backward derivative at t yields

$$\bar{\dot{A}} = -\bar{\dot{x}}_1 I(T)/x_1^2. \tag{6}$$

Using Equation 6 in conjunction with Equations 2 and 5 yields Equation 3.

Dynamics of x_2 can be found by directly quantifying the mean field behavior of $\mathbf{X}_2$. Consider the size of a cluster to which a focal individual, a sampled infected at time T, belongs. As before, $p_c \times f_{SI}$ gives the rate of coalescence. Two clusters merge at each coalescent event, so there is a probability proportional to $2/A$ that a focal individual belongs to a cluster that takes part in the event. And given that the individual's cluster coalesces, the average amount by which the cluster increases is x_1. Multiplying these factors and probabilities together yields

$$\bar{\dot{x}}_2 = p_c f_{SI}\frac{2}{A}x_1 = 2\bar{\dot{x}}_1. \tag{7}$$

As with x_1, this can be solved by integrating in reverse time with initial conditions $x_2(T;\ T) = 1$.

The variance of $\mathbf{X}_1$ can be found by noting that

$$\begin{aligned} E(\mathbf{X}_1^2) &= \sum_i i^2 \Pr\{\mathbf{X}_1 = i\} \\ &= \left(\sum_i i \Pr\{\mathbf{X}_1 = i\}\right)\left(\frac{\sum_i i^2 \Pr\{\mathbf{X}_1 = i\}}{\sum_i i \Pr\{\mathbf{X}_1 = i\}}\right). \end{aligned} \tag{8}$$

Recall that $\mathbf{X}_2$ is the size of a cluster selected with probability proportional to size, so

$$\Pr\{\mathbf{X}_2 = i\} = i \Pr\{\mathbf{X}_1 = i\}/\sum_j j \Pr\{\mathbf{X}_1 = j\}.$$

Combining the last two expressions with the definition of $x_1 = \sum_i i \Pr\{\mathbf{X}_1 = i\}$ gives

$$E(\mathbf{X}_1^2) = x_1 x_2.$$

Then, the variance in cluster size is

$$\operatorname{Var}(\mathbf{X}_1) = E(\mathbf{X}_1^2) - (E(\mathbf{X}_1))^2 = x_1 x_2 - x_1^2. \tag{9}$$

Higher moments can also be derived recursively from earlier moments. We now show that the nth moment of the CSD, M_n, can be found by solving the following differential equation with initial conditions $M_n(T) = 1$,

$$\bar{\dot{M}}_n = f_{SI}\frac{A}{I^2}\sum_{i=0}^{n-1}\binom{n}{i}M_i M_{n-i}, \tag{10}$$

where we define $M_0 := 1$ for convenience. Equations 3 and 4 could be derived using Equation 10 as a starting point.

Equation 10 is obtained by multiplying the rate at which a cluster merges with other clusters ($f_{SI}A/I^2$) and the expected change in the nth moment when two clusters merge. When a cluster of size i merges with a cluster of size j, the nth moment to be considered will change from that for a cluster of size i to that for a cluster of size $(i + j)$. To find the expected change in the nth moment when two clusters merge, we sum over all possible combinations of clusters of sizes i and j:

$$\begin{aligned} &\sum_i \sum_j \Pr\{\mathbf{X}_1 = i\}\Pr\{\mathbf{X}_1 = j\}(i+j)^n - i^n \\ &\quad = -M_n + \sum_i \Pr\{\mathbf{X}_1 = i\}\sum_j \Pr\{\mathbf{X}_1 = j\}\sum_{m=0}^{n}\binom{n}{m}i^{n-m}j^m \\ &\quad = -M_n + \sum_i \Pr\{\mathbf{X}_1 = i\}\sum_{m=0}^{n}\binom{n}{m}i^{n-m}\sum_j \Pr\{\mathbf{X}_1 = j\}j^m \\ &\quad = -M_n + \sum_i \Pr\{\mathbf{X}_1 = i\}\sum_{m=0}^{n}\binom{n}{m}i^{n-m}M_m \\ &\quad = -M_n + \sum_{m=0}^{n}\binom{n}{m}M_{n-m}M_m \\ &\quad = \sum_{m=0}^{n-1}\binom{n}{m}M_{n-m}M_m. \end{aligned}$$

The product of the coalescent rate $f_{SI}A^2/I^2$ and the factor $1/A$ that accounts for the probability that a focal lineage takes part in a coalescent event, along with the expected size function, yields Equation 10. In supporting information, Figure S1, we compare solutions of this equation to the second through fifth moments from simulations.

Fitting epidemic models to sequence data: If we know the branching times $t_1, t_2, \cdots, t_{n-1}$ for a phylogeny constructed from n sequences, we can use Equation 2 to fit an SIR model. In practice, there is considerable uncertainty about the exact genealogy and branching times given a sample of sequences. The theory developed here is based on a fixed genealogy with no uncertainty about branch lengths, but it should be straightforward to generalize these results to cope with error in the t_i (Drummond *et al.* 2005).

The total number of coalescent events observed between times t and T is proportional to $A(T;\ T) - A(t;\ T)$, and at some time $t < \tau < T$, the fraction of the coalescent events that have occurred is

$$F(\tau) = \frac{A(T;\ T) - A(\tau;\ T)}{A(T;\ T) - A(t;\ T)}. \tag{11}$$

This provides a cumulative distribution function for the distribution of coalescent times. Differentiating with respect to τ yields the density

$$-\bar{\dot{A}}/(A(T;\ T) - A(t;\ T)).$$

We make the approximation that when two lineages coalesce, the rates at which other lineages coalesce remain unchanged. Then each coalescent time will be an i.i.d. random variable with the distribution (11). The probability of observing a particular sequence of

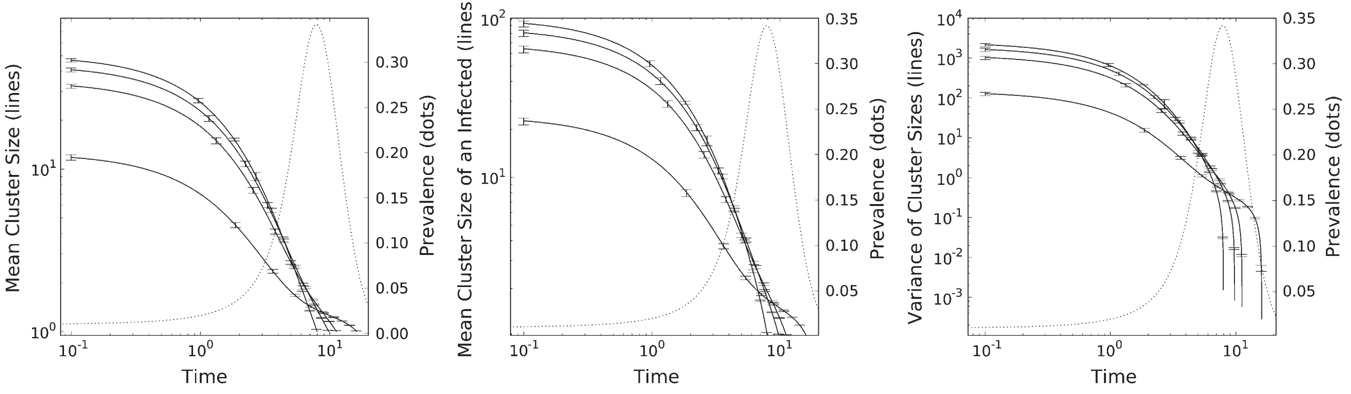


FIGURE 2.—The moments of the cluster size distribution over time as calculated by Equations 3 and 9 (lines, log scale). Four trajectories of the cluster size moments were generated for 4 sample times T. And for each trajectory, simulated moments were calculated for 10 threshold times t. Error bars show the 90% interval for 100 agent-based simulations [$N = 10^5$ and $I(0) = 1\%$]. The SIR model is $\dot{S} = -\beta SI$, $\dot{I} = \beta SI - \gamma I$, $\dot{R} = \gamma I$. Epidemic prevalence (dotted line) is shown on the right axis. Transmission rate $\beta = 1$, and recovery rate $\mu = 0.3$.

branching times will be proportional to the product of the density evaluated at each branching time. Consequently, we can construct the log-likelihood function out of an A-trajectory

$$\begin{aligned}\Lambda(t_1, \cdots, t_{n-1} \mid \theta) \\ &= \sum_{i=1}^{n-1} \log(-\dot{A}(t_i)/(A(T) - A(t))) \\ &= -(n-1)\log(A(T;\, T) - A(t;\, T)) + \sum_{i=1}^{n-1} \log(-\dot{A}(t_i;\, T)),\end{aligned} \tag{12}$$

where θ denotes the parameters of the SIR model, such as transmission and recovery rates. In File S1 we also present a fitting criterion based on the Kolmogorov–Smirnov statistic for comparing distributions.

RESULTS

Equation 3 indicates some simple relationships that govern coalescent rates in epidemics. Coalescent rates are proportional to epidemic incidence (f_{SI}) and inversely proportional to square prevalence (I^{-2}). Rates will be highest when prevalence is low and incidence is high, such as at the beginning of an epidemic, during the expansion phase, or following a trough in prevalence.

Equation 9 implies that variance of the CSD asymptotically approaches the mean squared (Figure S4). This is similar to what is seen in the offspring distribution of forward time branching processes, such as the Galton–Watson process (ATHREYA and NEY 2004).

The point in time where the ancestor function (5) crosses the value $1/N$ is the point at which the phylogeny of the virus has collapsed to a single lineage—the MRCA of the sequences. Therefore, if we collect a sample of size n at time T, and solve Equation 2 to time zero, with $A(T) = n/N$, the time τ that satisfies $A(\tau) = 1/N$ corresponds to the time to the most recent common ancestor of the sample. Although our differential equations should not serve as an adequate description of the discrete valued processes for values close to $1/N$, we find that this approximation works quite well. A demonstration with comparison to simulations is provided in Figure S11.

Simulations: To assess the performance of our model, we carried out stochastic simulations of SIR epidemics. Simulations were individual based and in continuous time. Transmission events and recovery events were queued using exponentially distributed lag times, similar to the Gillespie algorithm. Each transmission event was recorded, which allowed us to simulate viral phylogenies under controlled conditions and to test the accuracy of Equations 3 and 9. The transmission data were then converted into phylogenetic trees with known branching times.

Simulation code was independently written by S. D. Frost and E. M. Volz in Python and C. Results from both models were compared to ensure accuracy.

To assess the accuracy of the equations we have derived, we developed a simulation experiment with 10^3 (1%) initially infected agents out of a population of total size $N = 10^5$ otherwise identical agents. Transmission and recovery rates were such that $R_0 = 10/3$. Figure 2 shows Equations 3 and 9 (lines) and the 90% confidence intervals from simulations at 10 thresholds (t values). The exact values of t and T are reported in File S1. Each trajectory corresponds to a cross-sectional census of the infected population at four time points (T values) corresponding to maximum prevalence, as well as 86, 68, and 22% of maximum prevalence after the peak. As we go backward in time, all moments of the CSD increase, until the entire census of infecteds falls into a single cluster. Many of the trajectories intersect, which demonstrates that the CSD is a complex function of both t and T and could therefore not be reduced to a simple forward-looking model.

Comparison with the generalized skyline: Further simulations were developed to test the suitability of the

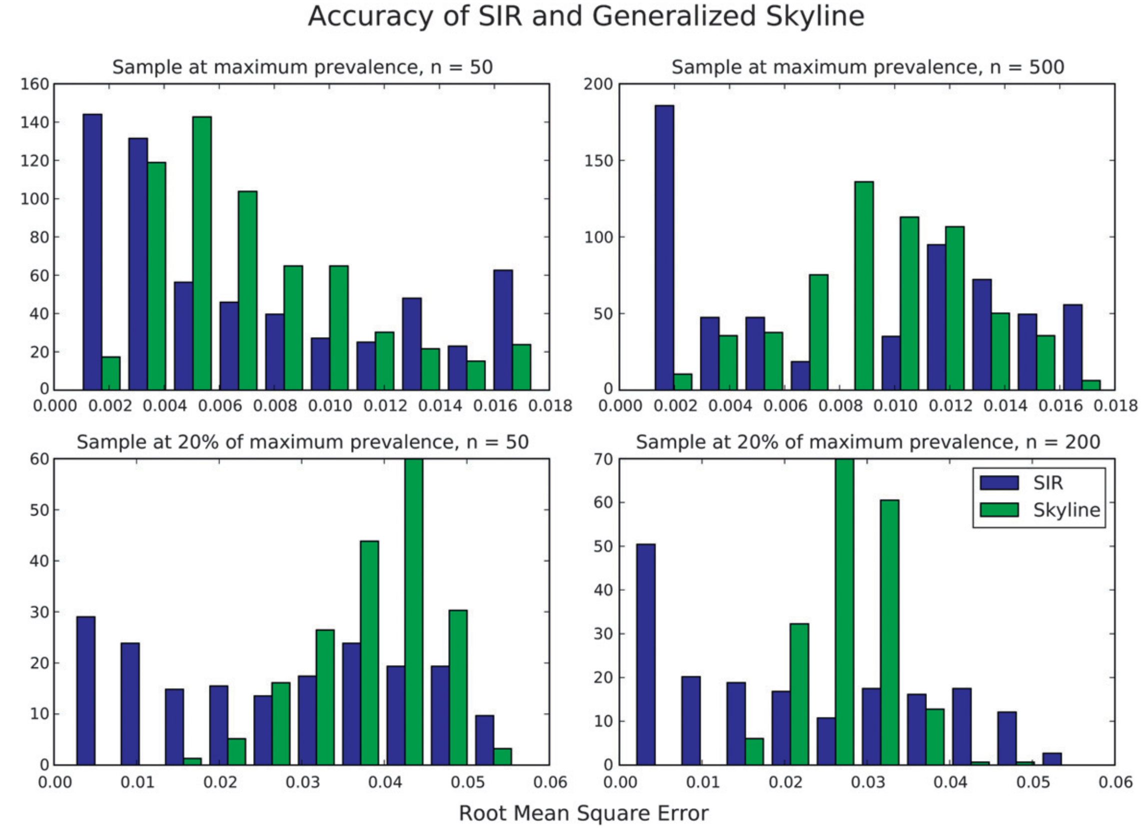


FIGURE 3.—Root mean square error of SIR and generalized skyline estimates of epidemic prevalence. Data are based on 300 simulated epidemics ($R_0 = 2$). RMSE is averaged over 100 time points.

model for estimating epidemiological parameters. When the number of infecteds is small, epidemic dynamics will be subject to large stochastic fluctuations. To determine if Equation 12 can be used to fit SIR models when the population size is small, we conducted a set of simulations with only a single initial infected in a population of 10,000 agents (Figure S5).

The simulations were also designed to determine if SIR models that are fit via likelihood Equation 12 can provide advantages beyond methods that are commonly used to estimate effective population size (N_e). For purposes of comparison, we used the generalized skyline model (OPGEN-RHEIN *et al.* 2005) (ape library in R) and compared the estimated effective population size to the best-fit SIR models and the known epidemic prevalence from simulations. Details of the simulations are provided in File S1.

We found that the accuracy of the best-fit SIR models exceeded that of the generalized skyline by 8–30% as measured by the root mean square error (RMSE) of estimated prevalence. It may seem surprising that the SIR model based on ODEs outperforms the generalized skyline even in the presence of stochasticity at small population sizes. This is due to the fact that population dynamics converge to the deterministic SIR model as the infected population increases in size. Fluctuating incidence due to stochastic effects when the number of infecteds is small has the effect of shifting the distribution of coalescence times to the left or the right, but does not fundamentally change the shape of the distribution. This is easily accounted for by including a parameter that varies the fraction initially infected in the deterministic SIR model.

Figure 3 shows the distribution of RMSE over 300 simulations. The mode of RMSE for the SIR model is zero for all experiments, whereas the skyline is slightly biased. Increasing sample size decreases RMSE for both SIR and skyline. Taking the sample at a later time (corresponding to 20% of peak prevalence) decreases the accuracy of both SIR and skyline, although in general the SIR models cope better with late sample times than does the skyline. In Figure S10, we show several representative SIR and skyline fits. It is usually the case that the generalized skyline fails to detect a decrease in prevalence and overestimates in the latter stages of the epidemic.

The SIR models also provide a quite accurate estimate of R_0 [$R_0 = 2$, $\hat{R}_0 = 1.95$ (95%: 1.71–2.17)].

The effect of a sample fraction: In the Kingman coalescent, the fraction of the population that is sampled is assumed to be small, such that the probability that more than two individuals have the same parent in the preceding generation is negligible. For example,

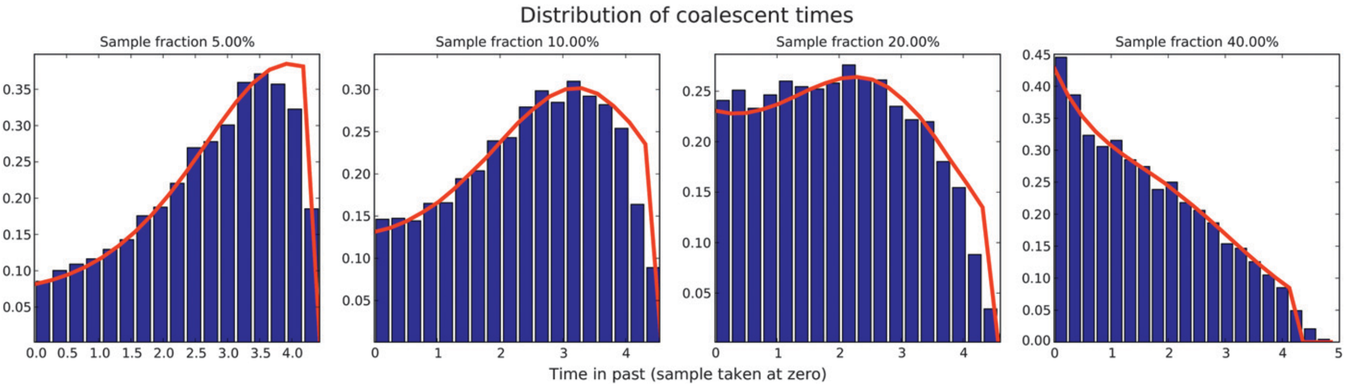


FIGURE 4.—The empirical distribution of coalescence times based on 150 simulated SIR epidemics. Transmission rate = 2, recovery rate = 1. The expected distribution based on Equation 11 is shown in red.

Kingman showed that the probability that n sampled sequences will not have a common ancestor in the preceding generation is

$$\prod_{i<n}(1-i/N)$$
$$= 1-\sum_{i<n}\frac{i}{N}+O(N^{-2}) = 1-\binom{n}{2}/N+O(N^{-2}).$$

Kingman then made the approximation that the $O(N^{-2})$ terms are zero, which yields a minimum requirement that $n < \sqrt{2N}$.

Analytical work has been carried out to investigate the effect on coalescent processes of violating the assumption of a small sample fraction (see, for example, FU 2006), using discrete mathematics similar to the original Kingman model. Such work has indicated that the Kingman coalescent can be a surprisingly good approximation even when the sample fraction is large.

Nevertheless, our model is not an approximation and takes the sample fraction into account. This gives some insight into how the fraction of the infected population sampled affects the distribution of coalescent times and thus the shape of the reconstructed phylogeny of viral sequences.

Figure 4 shows the empirical distribution of coalescence times for 150 simulations ($R_0 = 2$) with samples taken at peak prevalence. The sample fraction was varied from 5 to 40%. When the sample fraction is small (5%), the distribution is skewed left, meaning the phylogeny is starlike, which is in agreement with conventional notions for an exponentially growing population. However, as the sample fraction is increased to 10, 20, and 40%, the shape of the distribution changes until it is skewed right, which means that most of the branches occur close to the tips. These qualitatively antipodal distributions are generated by the same underlying population dynamics, with only the sample fraction being varied. This observation is of practical as well as theoretical interest, since many serological surveys for HIV may reach >20% of infected individuals within a given locality (LEWIS *et al.* 2008).

Equation 11 gives the analytical distribution of coalescence times and is shown in red in Figure 4. It also provides some simple intuition for why most coalescence events will happen close to the sample time (T) when the sample fraction is large. We use the initial conditions $A(T) = n/N$, so that when n is large, the term $(A(T)/I(T))^2$ is also large, which is the probability that two individuals in a transmission event represent sample lineages. Conversely, if n and $(A(T)/I(T))^2$ are small, fewer coalescent events will occur until I converges to A, which will occur early in the epidemic.

Estimating HIV prevalence: Equation 2 gives the rate of coalescence at any time prior to the sample time (T) and, by extension, the distribution of coalescence times. This allowed us to derive the likelihood function (12), which we used to fit a simple mass-action SIR model to 55 HIV-1 sequences of the *pol* gene collected as part of the ACTG241 clinical trial (D'AQUILA *et al.* 1996; LEIGH BROWN *et al.* 1999). All sequences were collected from men who have sex with men (MSM) over a short period of time (May to July, 1993) within the United States. Because the sequences were collected within a short window of time, it is valid to make the approximation that all sequences were sampled simultaneously. To estimate a phylogeny, we used a general-time-reversible model of nucleotide substitution (TAVARE 1986) with gamma-distributed variation in site-to-site substitution rates. The root giving the most clocklike rates was determined by maximum likelihood and the null hypothesis of a molecular clock could not be rejected at the 5% significance level.

The epidemiology of HIV has several factors that are important to include in a model. Upon infection, individuals progress through an acute phase lasting 1–3 months and then progress to a chronic phase lasting many years. The transmission probability per act is much greater during the acute phase. Furthermore, since we wish to model the epidemic over a period of 25 years, we must consider natural mortality and immigration into the susceptible pool. All of these factors are considered in the following model:

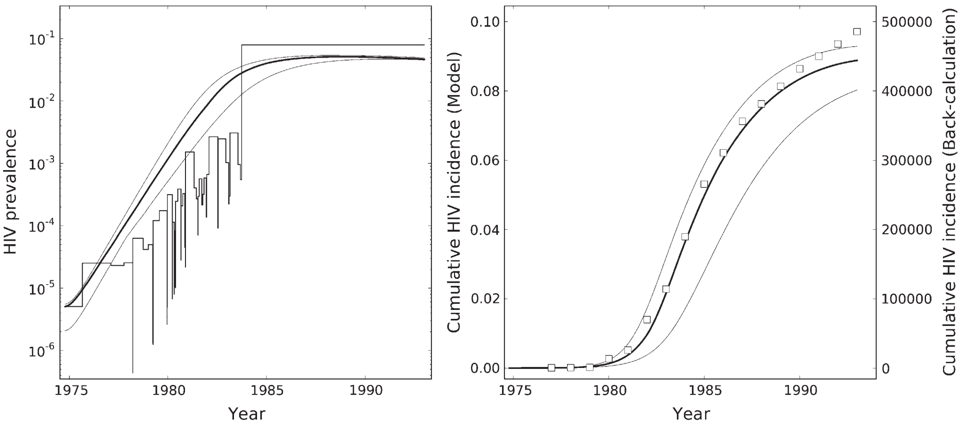


FIGURE 5.—Left: Estimated epidemic prevalence (logarithmic scale) of HIV among MSM in the United States. A solution to Equation 16 is compared to the skyline plot, rescaled such that minimum effective population size equals minimum prevalence. The thin lines show 95% confidence intervals. Right: Estimated cumulative incidence of HIV among MSM *vs.* time (years prior to 1993). A solution to Equation 16 is compared to estimates based on sero-surveillance data (HALL *et al.* 2008).

$$\dot{S} = -S^{\alpha}(\beta_1 I_1 - \beta_2 I_2) + \mu - \mu S \qquad (13)$$

$$\dot{I_1} = S^{\alpha}(\beta_1 I_1 + \beta_2 I_2) - \gamma_1 I_1 - \mu I_1 \qquad (14)$$

$$\dot{I_2} = \gamma_1 I_1 - \gamma_2 I_2 - \mu I_2. \qquad (15)$$

I_1 and I_2 respectively represent the fractions of the population that are at the acute and the chronic stages of infection. Parameters we wish to estimate include the following:

β_1: The transmission rate of acute infecteds.

β_2: The transmission rate of chronic infecteds.

μ: The immigration rate into the susceptible population and the natural mortality rate. We consider immigration to balance natural mortality.

α: A parameter that controls how incidence scales with cumulative incidence.

ε: The fraction of the population infected at the TMRCA of the sample.

Many more parameters could be included in a model for HIV among MSM, but since our purpose is to fit a model to only 55 sequences, we choose to keep the number of free parameters to a minimum. In addition, we assumed an acute phase that lasts 2 months on average ($\gamma_1 = 1/60$) and a chronic phase that lasts 10 years on average [$\gamma_2 = 1/(10 \times 365)$].

Prior distributions are given in File S1.

Given $n = 55$ sequences, we use the initial conditions $A(T) = 55/N$, $I_1(0) = \varepsilon$, and $S(0) = 1 - \varepsilon$. Since we are including equations for two types of infecteds, we must keep track of ancestor functions for both types. A_1 and A_2 are the fractions of the population that are respectively acute and chronic infected and that have sampled progeny at time T. We have

$$\dot{\bar{A_2}} = -\gamma_1 I_1(A_2/I_2) + \beta_2 I_2 S^{\alpha}(A_1/I_1)((I_2 - A_2)/I_2) \qquad (16)$$

$$\dot{\bar{A_1}} = \gamma_1 I_1(A_2/I_2) - \beta_1 I_1 S^{\alpha}(A_1/I_1)^2 - \beta_2 I_2 S^{\alpha}(A_1/I_1). \qquad (17)$$

For purposes of fitting the SIR model, we use $A = A_1 + A_2$ and $\dot{\bar{A}} = \dot{\bar{A_1}} + \dot{\bar{A_2}}$. A derivation is provided in File S1.

Fitting the model proceeded in two steps. First, we fit a model using Equation 12 as described above. The second step made use of sero-surveillance data of MSM in the United States (HALL *et al.* 2008). These data provided estimates of HIV incidence based on back calculation for the period 1977–2006. To ameliorate error from uncertainty in the chronological values of phylogenetic branch lengths, we adjusted the timescale of the epidemic and rescaled estimated rates to gain the greatest fit with incidence data by a least-squares criterion.

Figure 5 shows the best-fit SIR model. The median posterior estimates were as follows: acute transmission rate, $\hat{\beta}_1 = 1$ transmission per 47 days; chronic transmission rate, $\hat{\beta}_2 =: 1$ transmission per 1207 days; immigration rate to susceptible state, $\hat{\mu} = 1$ per 19.5 years; and incidence scaling parameter, $\hat{\alpha} = 9.77$. Together, these parameters imply an R_0 value of 2.24 (see File S1). They also imply that 41% of transmissions occur during the acute stage.

For comparison with our SIR model, effective population size (N_e) was calculated using the skyline plot (PYBUS *et al.* 2000). N_e was rescaled so that $\min(N_e) = \min(I)$. Figure 5 shows the rescaled skyline and an SIR trajectory that was integrated with parameters from the median of the posterior distribution. Confidence intervals are also given, which show the upper and lower bounds within which 95% of posterior epidemic prevalence falls. Figure 5 also compares the best-fit SIR model with the estimated cumulative incidence among MSM in the United States based on sero-surveillance data. The SIR model is in broad agreement with the data from public health sources regarding the early rate of growth and saturation in the early 1990s. The skyline also reproduces the growth rate during the expansion phase and the tapering of epidemic growth in the early 1990s. However, the skyline predicts a rise in N_e between 1980 and 1993, which probably overestimates the true prevalence.

We have also compared the CSD mean and variance from our best-fit SIR model to the empirical values from the ACTG241 data (Figure 6). The SIR model successfully reproduces the mean cluster size throughout the

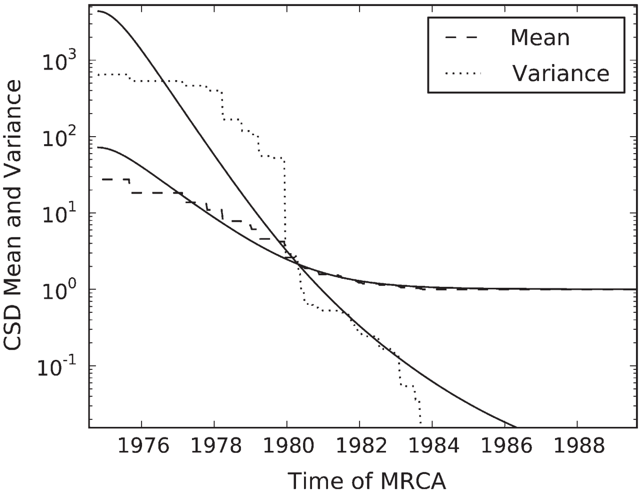


FIGURE 6.—The mean cluster size (dashes) and variance of cluster sizes (dotted line) are calculated from the empirical observations from the ACTG241 sequences (dashed lines) and compared to our best-fit SIR model (solid lines). The horizontal axis gives the clustering threshold as the year of the MRCA of a cluster.

course of the epidemic. However, there is substantial deviation between the actual and the predicted variance of cluster sizes. As the clustering threshold is increased, all sampled infecteds eventually fall within a single cluster, and in a finite population, variance converges to zero (not shown).

DISCUSSION

The distribution of cluster sizes is a function of the time T at which we observe a population, such as by taking a sample of sequences, and $t < T$, which is a clustering threshold (if the MRCA of two sequences occurs after t, then those sequences are clustered). We have derived differential equations that describe how the moments of the CSD change as the threshold t moves into the past. This could be used to calculate the distribution of cluster sizes to arbitrary precision at any time. It is straightforward to use the model to calculate the probability that an infected host will have viral progeny at a later time point and, conversely, the expected number of ancestor lineages of a sample taken at T. The model promises to serve as a null hypothesis for clustering of infecteds under various epidemiological scenarios and could possibly be used to detect effects that may distort the CSD such as selection and population structure.

The CSD is sensitive to details of the underlying population dynamics. Most coalescent approaches take into account only variable population size, such as epidemic prevalence, but not variable birth rates, analogous to epidemic incidence. Such approaches can give misleading results for epidemics. For example, in both susceptible–infected (SI) models (no recovery) and susceptible–infected–susceptible (SIS) models (recovery into the susceptible state), prevalence rapidly approaches an equilibrium. However, a naive coalescent model based on constant population size would erroneously predict identical coalescent patterns in these two cases. In fact, the SIS case is very similar to a standard constant-population size coalescent, but the lineages in an SI epidemic coalesce only during exponential growth, not at equilibrium (Figure S2 and Figure S3).

We observed drastically less precision when estimating recovery rates than when estimating transmission rates. Consequently, decline in prevalence is much harder to detect than growth. This has been observed previously (LAVERY *et al.* 1996) in other biological systems due to differences in the timescale of population change and genetic variation. We nevertheless found that our estimation procedure is robust to misspecification of priors that include zero recovery, and it is feasible to distinguish SI from SIR dynamics (Figure S6, Figure S7, Figure S8, and Figure S9).

In conclusion, coalescent-based estimates of effective population size, such as the generalized skyline, have wide applicability and require minimal consideration of underlying population dynamics. However, in the case that the epidemic dynamics are well understood, the potential is raised for a population genetic model that takes into account the precise effects of transmission and recovery, thereby predicting population dynamics with greater accuracy. We have developed a model that provides a step toward the formal integration of phylodynamics and epidemiology and that can be used to estimate epidemiological and demographic parameters directly from viral sequence data.

Fitting population models to data requires biological simplifications to make the model tractable, which presents the danger of making the model useless for real systems (WILSON *et al.* 2005). Pathogens require successful reproduction both within and between hosts, whereas we have focused entirely on transmission of lineages to uninfected and immunologically naive hosts. We have not considered biological nuances such as superinfection and recombination or the possibility that different strains will have different epidemiological characteristics. Consequently, there are many ways that our model could be extended and improved.

We have calculated coalescent rates and CSD moments only for the most simple mass-action SIR models. But modern mathematical epidemiology has progressed in the direction of incorporating variable host susceptibility, pathogen virulence, geographical heterogeneity, and host contact network structure. Reproducing our derivations for such models would be a difficult but worthy enterprise.

While we have focused on variable population size in epidemics, a second pillar of phylodynamics concerns the effects of immune selection on viral phylogenies (GRENFELL *et al.* 2004). A major limitation of our approach is that we adopt the standard assumption of selective neutrality. It is unknown how our method would perform for genes under strong immune selection, such as influenza virus hemagglutinin.

We have made a first attempt at a method for fitting arbitrary SIR models to cross-sectional samples of viral sequences. Many challenges remain for increasing the utility of the method. It may be possible to improve estimation of model parameters when historical prevalence data are available. However, it is not known how to discriminate between competing models when only sequence data are available. The estimation theory developed here is based on a fixed genealogy of virus with no uncertainty about branch lengths; in reality there can be a great deal of uncertainty about the structure of the genealogy, and it should be straightforward to generalize the method to account for this (Drummond *et al.* 2005). Finally, it should also be possible to extend our solutions to heterochronous samples—sequence data collected at multiple time points over the course of an epidemic.

Irene Hall provided estimates of HIV incidence in MSM. The authors acknowledge support from the National Institutes of Health (T32 AI07384, R01 AI47745). S.D.W.F. is supported by a Royal Society Wolfson Research Merit Award. M.J.W. is supported by the Biotechnology and Biological Sciences Research Council.

LITERATURE CITED

Anderson, R. M., and R. M. May, 1991 *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, London/New York/Oxford.

Athreya, K. B., and P. E. Ney, 2004 *Branching Processes*. Dover, New York.

Bailey, N. T. J., 1975 *The Mathematical Theory of Infectious Diseases and Its Applications*. Hafner Press, New York.

Brenner, B. G., M. Roger, J. Routy, D. Moisi, M. Ntemgwa *et al.*, 2007 High rates of forward transmission events after acute/early HIV-1 infection. J. Infect. Dis. **195:** 951.

Brown, A. J., H. F. Günthard, J. K. Wong, R. T. D'Aquila, V. A. Johnson *et al.*, 1999 Sequence clusters in human immunodeficiency virus type 1 reverse transcriptase are associated with subsequent virological response to antiretroviral therapy. J. Infect. Dis. **180:** 1043–1049.

D'Aquila, R. T., M. D. Hughes, V. A. Johnson, M. A. Fischl, J. P. Sommadossi *et al.*, 1996 Nevirapine, zidovudine, and didanosine compared with zidovudine and didanosine in patients with HIV-1 infection: a randomized, double-blind, placebo-controlled trial. Ann. Intern. Med. **124:** 1019–1030.

Drummond, A. J., A. Rambaut, B. Shapiro and O. G. Pybus, 2005 Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. **22:** 1185–1192.

Drumright, L. N., and S. D. W. Frost, 2008 Sexual networks and the transmission of drug-resistant HIV. Curr. Opin. Infect. Dis. **21:** 644.

Fu, Y., 2006 Exact coalescent for the Wright–Fisher model. Theor. Popul. Biol. **69:** 385–394.

Goodreau, S. M., 2006 Assessing the effects of human mixing patterns on HIV-1 interhost phylogenetics through social network simulation. Genetics **172:** 2033–2045.

Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly *et al.*, 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. Science **303:** 327.

Griffiths, R. C., and S. Tavare, 1994 Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. B Biol. Sci. **344:** 403–410.

Hall, H., R. Song, P. Rhodes, J. Prejean, Q. An *et al.*, 2008 Estimation of HIV incidence in the United States. J. Am. Med. Assoc. **300:** 520.

Hue, S., D. Pillay, J. P. Clewley and O. G. Pybus, 2005 Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc. Natl. Acad. Sci. USA **102:** 4425–4429.

Kaj, I., and S. M. Krone, 2003 The coalescent process in a population with stochastically varying size. J. Appl. Probab. **40:** 33–48.

Kermack, W. O., and A. G. McKendrick, 1927 A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. Ser. A, *Containing Papers of a Mathematical and Physical Character,* **115:** 700–721.

Kingman, J. F. C., 1982a On the genealogy of large populations. J. Appl. Probab. **19:** 27–43.

Kingman, J. F. C., 1982b The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Lavery, S., C. Moritz and D. R. Fielder, 1996 Genetic patterns suggest exponential population growth in a declining species. Mol. Biol. Evol. **13:** 1106–1113.

Lewis, F., G. J. Hughes, A. Rambaut, A. Pozniak and A. J. Leigh Brown, 2008 Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med. **5:** e50.

Nee, S., E. C. Holmes, A. Rambaut and P. H. Harvey, 1996 Inferring population history from molecular phylogenies, pp. 66–80 in *New Uses for New Phylogenies*, edited by P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith and S. Nee. Oxford University Press, Oxford.

Opgen-Rhein, R., L. Fahrmeir and K. Strimmer, 2005 Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. BMC Evol. Biol. **5:** 6.

Pao, D., M. Fisher, S. Hué, G. Dean, G. Murphy *et al.*, 2005 Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. AIDS **19:** 85.

Pybus, O. G., A. Rambaut and P. H. Harvey, 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics **155:** 1429–1437.

Pybus, O. G., M. A. Charleston, S. Gupta, A. Rambaut, E. C. Holmes *et al.*, 2001 The epidemic behavior of the hepatitis C virus. Science **292:** 2323–2325.

Robbins, K. E., P. Lemey, O. G. Pybus, H. W. Jaffe, A. S. Youngpairoj *et al.*, 2003 US human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. J. Virol. **77:** 6359–6366.

Rosenberg, N. A., and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. **3:** 380–390.

Tavare, S., 1986 Some probabilistic and statistical problems in the analysis of DNA sequences, pp. 57–86 in *Lectures on Mathematics in the Life Sciences*. American Mathematical Society, Providence, RI.

Wilson, D. J., D. Falush and G. McVean, 2005 Germs, genomes and genealogies. Trends Ecol. Evol. **20:** 39–45.

Yerly, S., S. Vora, P. Rizzardi, J. P. Chave, P. L. Vernazza *et al.*, 2001 Acute HIV infection: impact on the spread of HIV and transmission of drug resistance. AIDS **15:** 2287.

Yusim, K., M. Peeters, O. G. Pybus, T. Bhattacharya and B. Korber, 2001 Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. Philos. Trans. R. Soc. B Biol. Sci. **356:** 855–866.

Communicating editor: M. W. Feldman

Simple Epidemiological Dynamics Explain Phylogenetic Clustering of HIV from Patients with Recent Infection

Erik M. Volz[1]*, James S. Koopman[1], Melissa J. Ward[2], Andrew Leigh Brown[2], Simon D. W. Frost[3]

1 Department of Epidemiology, University of Michigan, Ann Arbor, Michigan, United States of America, **2** Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom, **3** Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom

Abstract

Phylogenies of highly genetically variable viruses such as HIV-1 are potentially informative of epidemiological dynamics. Several studies have demonstrated the presence of clusters of highly related HIV-1 sequences, particularly among recently HIV-infected individuals, which have been used to argue for a high transmission rate during acute infection. Using a large set of HIV-1 subtype B pol sequences collected from men who have sex with men, we demonstrate that virus from recent infections tend to be phylogenetically clustered at a greater rate than virus from patients with chronic infection ('excess clustering') and also tend to cluster with other recent HIV infections rather than chronic, established infections ('excess co-clustering'), consistent with previous reports. To determine the role that a higher infectivity during acute infection may play in excess clustering and co-clustering, we developed a simple model of HIV infection that incorporates an early period of intensified transmission, and explicitly considers the dynamics of phylogenetic clusters alongside the dynamics of acute and chronic infected cases. We explored the potential for clustering statistics to be used for inference of acute stage transmission rates and found that no single statistic explains very much variance in parameters controlling acute stage transmission rates. We demonstrate that high transmission rates during the acute stage is not the main cause of excess clustering of virus from patients with early/acute infection compared to chronic infection, which may simply reflect the shorter time since transmission in acute infection. Higher transmission during acute infection can result in excess co-clustering of sequences, while the extent of clustering observed is most sensitive to the fraction of infections sampled.

Citation: Volz EM, Koopman JS, Ward MJ, Brown AL, Frost SDW (2012) Simple Epidemiological Dynamics Explain Phylogenetic Clustering of HIV from Patients with Recent Infection. PLoS Comput Biol 8(6): e1002552. doi:10.1371/journal.pcbi.1002552

Editor: Christophe Fraser, Imperial College London, United Kingdom

Received September 27, 2011; **Accepted** April 24, 2012; **Published** June 28, 2012

Funding: This study was funded by the NIH grant no. 1-K01-AI-091440-01 and NIH grant no. R01-AI078752. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: erikvolz@umich.edu

Introduction

Phylogenetic clusters of closely related virus such as HIV arise from the epidemiological dynamics and transmission by infected hosts. If virus is phylogenetically clustered, it is an indication that the hosts are connected by a short chain of transmissions [1].

If super-infection is rare, and assuming an extreme bottleneck at the point of transmission, each lineage in a phylogenetic tree corresponds to a single infected individual with its own unique viral population [2,3]. A transmission event between hosts causes an extreme bottleneck in the population of virus in the new hosts. For infections between MSM, it is estimated that infection is initiated by one or several virions [4,5]. At the time of transmission, the quasispecies of virus within the transmitting host diverges and can thereby generate a new branch in the phylogeny of consensus viral isolates from infected individuals [6]. Transmissions in the recent past should be reflected by recently diverged lineages, and transmissions from long ago should reflect branches close to the root of a tree. [7]. Viruses such as HIV which have a high mutation rate relative to epidemiological spread can generate epidemics such that the correspondence between transmission and phylogenetic branching is most clear [2].

Given a phylogeny of virus reconstructed from n samples, the phylogenetic clusters are a partition of the n sample units into disjoint sets as a function of the tree topology. A cluster will consist of all taxa of the tree that are descended from a given lineage on the interior of the tree. There are many variations of this idea, and there is no general agreement about how to choose interior lineages for defining clusters. The most common algorithms require strong statistical support for a monophyletic clade among all taxa in a cluster [8–14]. These definitions may additionally require all taxa in a cluster to be connected by short branches with less than a threshold length [11], or similarly require that the genetic sequences corresponding to each taxon be separated by a genetic distance less than a given threshold [8,14]. Definitions of clustering based on statistical support for monophyly are very difficult to operationalize in a mathematical model, and in particular, it is not clear how the statistical significance of internal nodes relates to population dynamics. Consequently, we have devised a conceptually similar definition of clusters that relies on the estimated time to most recent common ancestor (TMRCA) of a set of taxa [15]. A formal definition is provided below.

The sizes of the groupings that arise from a clustering algorithm have been interpreted as a reflection of the heterogeneity of epidemiological transmission. The distribution of cluster sizes of HIV is often skewed right, and depending on the definition of clustering used, can have a heavy tail [14,15]. This is consistent with the prevailing view among modelers of sexually transmitted infections that there is a skewed and in some cases power-law distribution in the number of risky sexual contacts in the

Author Summary

Diversity of viral genetic sequences depends on epidemiological mechanisms and dynamics, however the exact mechanisms responsible for patterns observed in phylogenies of HIV remain poorly understood. We observe that virus taken from patients with early/acute HIV infection are more likely to be closely related. By developing a mathematical model of HIV transmission, we show how these and other patterns arise as a simple consequence of intensified transmission during the early/acute stage of HIV infection, however observing these patterns is highly dependent on sampling a significant fraction of prevalent infections.

population, however it is not straightforward to make inferences about sexual network properties from cluster size distributions [16]. In the case of HIV, the distribution of branch lengths within clusters may also reflect the disproportionate impact of early and acute HIV infection on forward transmission, which is due to higher viral loads in the early stages of infection, higher transmissibility per act [17], and fluctuating risk behavior [18].

When the taxa of the phylogeny are labeled, such as with the demographic, behavioral or clinical attributes of the the individuals from whom the virus was sampled, one can further analyze statistical properties of clustered taxa. Similar taxa, such as those arising from acute infections, may cluster together (or *co-cluster*) at greater rates. Patterns of co-clustering might be informative about the fraction of transmissions that occur at different stages of infection or between different demographic categories. HIV phylogenies from men who have sex with men (MSM) have been widely observed [12,13,19] to have individuals with early/acute HIV infection that are much more likely to appear in a phylogenetic cluster. And moreover, if early-stage individuals are in a cluster, they are much more likely to be clustered with other early infections. Both Lewis et al. and Brenner et al. [8,9] have hypothesized that co-clustering of early infection is caused by higher transmissibility per act during early infection. For example, in phylogenies with time-scaled branch lengths, if a large fraction of clusters have a maximum branch length of six months [8,15], this suggests that *at least* that fraction of transmissions also occur within six months. In this article we demonstrate that the mechanisms that generate co-clustering of early infections are complex, and involve many attributes of the epidemic in addition to higher transmissibility per act [17]. To summarize, several features of the phylogenetic structure of HIV in MSM have been independently observed by several investigators:

- Many more early infections are phylogenetically clustered than late infections. For future reference, we will refer to this as *excess clustering* of early/acute infections.
- If an early infection is clustered, it is more likely to be co-clustered with another early infection than expected by chance alone. For future reference, we will refer to this as *excess co-clustering* of early/acute infections.
- The distribution of phylogenetic cluster sizes is skewed to the right and is potentially heavy-tailed.

Below, we illustrate these clustering patterns using 1235 HIV-1 subtype B *pol* sequences collected between 2004 and 2010 in Detroit, Michigan, USA.

These common clustering features motivate several questions. How informative are clustering patters about the underlying epidemic? In particular, how does higher transmissibility per act during early infection shape the phylogeny of virus ? To address these questions, we have developed a simple mathematical framework that demonstrates the connection between epidemiological dynamics and the expected patterns of clustering from a transmission tree and the corresponding phylogeny.

Our modeling work suggests that common features of HIV phylogenies are not coincidences, but universal features of certain viral phylogenies. We expect to see similar patterns for any disease such that the natural history features an early period of intensified transmission. High transmission rates during early infection may be a consequence of higher transmissibility per act due to high viral loads, but are also influenced by behavioral factors, such as fluctuating risk behavior [18], concurrency [20], and a lack of awareness of the infection. We do not explicitly model immunological or behavioral factors, but rather consider a compound parameter that describes the rate of transmission during the early/acute period. We find that while higher transmission rates increase the frequency of early/acute clustering, virus collected from early/acute patients clusters at a higher rate even when transmission rates are uniform over the infectious period.

Materials and Methods

Ethics statement

This research was reviewed by the Institutional Review Board at the University of Michigan. Data used in this research was originally collected for HIV surveillance purposes. Data were anonymized by staff at the Michigan Department of Community Health before being provided to investigators. Because this research falls under the original mandate for HIV surveillance, it was not classified as human subjects research.

Phylogenetic clustering of Michigan HIV-1 sequences

Our analysis consists of an empirical component which establishes clustering patterns for a geographically and temporally delineated set of HIV sequences, and an analytical component which establishes a possible mechanism that could generate the observed patterns.

We examined the phylogenetic relationships of 1235 HIV-1 subtype B partial-*pol* sequences originally collected for drug-resistance testing. All sequences were collected in the Detroit metropolitan statistical area between 2004 and 2010. Sequences were tested for quality and subtype using the LANL quality control tool [21–23], and aligned against a subtype-B reference (HXB2).Drug resistance sites [24] were treated as missing data.

A maximum clade credibility phylogeny was estimated with BEAST 1.6.2 [25]. The phylogeny was estimated using a relaxed molecular clock and and HKY85 model of nucleotide substitution with Gamma rate variation between sites (4 categories). The MCMC was run for 50 million iterations with sampling every 10^4 iterations. The first million iterations were discarded. The effective sample size of all parameters exceeded 50.

The phylogeny was converted into a matrix of pairwise distances between taxa expressed in units of calendar time. The distance between a pair of taxa was the TMRCA estimated by BEAST. Taxa were then classified into clusters using hierarchical clustering algorithms. A pair of taxa were considered to be clustered if the estimated TMRCA did not exceed a given threshold, and a range of thresholds was examined, from 0.5% of the maximum distance to the distance corresponding to the point where 90% of taxa are clustered with at least one other taxon.

Co-clustering of early/acute infections was investigated using a clinical variable (CD4 count) and a measure of genetic diversity of the virus. Both CD4 and sequence diversity are imprecise indicators of stage of infection. Nevertheless, with a large population-based sample, even noisy indicators of stage of infection are useful for illustrating phylodynamic patterns.

In most cases, CD4 counts were assessed contemporaneously with samples collected for sequencing. The CD4 cell counts can be informative about disease progression and can be used as a noisy predictor of the unknown date of infection [26]. Individuals with very high cell counts are unlikely to represent late/chronic infections, and we hypothesize that virus from these patients will be more likely to be phylogenetically clustered. Clustering of patients with high CD4 was previously observed by Pao et al. [10]

Recent work [27] has also highlighted the potential for sequence diversity to be informative of the date of infection. The frequency of ambiguous sites (FAS) in consensus sequences provides an approximate measure of sequence diversity in the host. HIV infection is initiated by one or a few founder lineages [4,5]; initially the diversity of the viral population within the host is low, but diversity increases steadily over the course of infection [28]. By convention, consensus sequences report ambiguous sites as those where the most frequent nucleotide is read with a frequency less than 80%. We hypothesize that having few ambiguous sites is an indicator of early/acute infection; sequences with fewer ambiguous sites will be more likely to be in a phylogenetic cluster and to be clustered with other sequences with few ambiguous sites.

A simple analysis was conducted to establish the existence of excess clustering and co-clustering in the Michigan sequences. This analysis is not designed to classify our sample into a early/acute component or to estimate the date of infection for each unit.

To illustrate excess clustering of early/acute infections, we calculated the mean CD4 cell count and FAS for each sample unit in a phylogenetic cluster. Because all clustering thresholds are arbitrary, we explored a large range of values, up to the point where 90% of the sample was clustered with at least one other unit. The standard error of the estimated mean was calculated assuming simple random sampling. For small threshold distances, very few taxa are clustered, and the standard error is large, but decreases monotonically as the threshold is increased and more taxa are clustered.

To illustrate excess co-clustering, we classified taxa into three categories of CD4: those with CD4 <200, representing AIDS cases; those with CD4 >800, and those with CD4 between 200 and 800. Taxa were also classified into quartiles by FAS. We then counted the number of pairwise clusterings of taxa within and between each category. These counts were arranged in a matrix. Large counts along the diagonal (within categories) represent co-clustering by stage of infection. To establish excess co-clustering, we compared the counts to the expectation if clusters were being formed at random, e.g. if two taxa were selected uniformly at random without replacement.We denote the symmetric matrix of co-clustering counts as M, so that M_{ij} represents the number of times that a taxon in category i is clustered with a taxon in category j. The sum of counts in the i'th row of M will be denoted m_i. Following the methods described in [29], the expected value of M under random pair formation is

$$\langle M_{ij} \rangle = m_i m_j / \sum_{ij} M_{ij}.$$

Below, we illustrate the difference $M_{ij} - \langle M_{ij} \rangle$. We can also calculate the assortativity coefficient [29], r, which describes the total amount of co-clustering in the matrix. To construct the co-clustering matrices, we selected the value of the distance threshold which maximized the assortativity coefficient.

Mathematical model

Following the approach outlined in [6] and [30], we develop a deterministic coalescent model derived from a compartmental susceptible-infected-recovered (SIR) model. A system of several ordinary differential equations describe the dynamics of prevalence of early and late HIV infection. Individuals pass from a susceptible state, to an early/acute infection state, to a chronic infection state followed by removal (treatment or death). S,I_1, and I_2 will denote the numbers susceptible, acute, and chronically infected respectively, and the population size will be denoted N. For didactic purposes, we will suppose that treatment is completely effective at preventing forward transmissions. The HIV model is described by the following equations:

$$\dot{S} = -\frac{S}{N}(\beta_1 I_1 + \beta_2 I_2)\theta(t) + b(t) - \mu S$$

$$\dot{I}_1 = \frac{S}{N}(\beta_1 I_1 + \beta_2 I_2)\theta(t) - \gamma_1 I_1 - \mu I_1 \qquad (1)$$

$$\dot{I}_2 = \gamma_1 I_1 - \gamma_2 I_2 - \mu I_2$$

In these equations, β_1 and β_2 are respectively the frequency-dependent transmission rates for early and chronic infected individuals. The average duration of early and chronic infection are respectively $1/\gamma_1$ and $1/\gamma_2$. Natural mortality occurs at the rate μ and immigration into the susceptible state occurs at the rate $b(t) = \mu(S + I_1 + I_2) + \gamma_2 I_2$, which maintains a constant population size $N = 10^4$. $\theta(t)$ is a term which modulates the way incidence of infection scales with prevalence. For the results presented below, we choose $\theta(t) = e^{-\alpha(I_1 + I_2)/N}$. This term corrects for observed patterns of decreasing incidence with prevalence; this can occur as a result of population heterogeneities (including sexual network structure) or as the result of decreasing risk behavior as knowledge of the epidemic spread. Many more relevant details could be included in a model of the HIV epidemic in MSM, however our purpose is to demonstrate how these simple dynamics lead to observed phylogenetic patterns.

In [6], a similar HIV model was presented along with a method to fit such models to a sequence of phylogenetic divergence times (the heights of nodes in a time-scaled phylogeny). Where possible, we will use the parameter estimates from [6]. The parameters are reported in table 1. Together, these parameters imply $R_0 = 2.24$ and that 41% of transmissions occur during the acute stage.

Corresponding to an epidemic model of the form 1, we can define a coalescent process [31,32] that describes the properties of the transmission tree and by extension the phylogeny of virus. The taxa descended from a lineage at time t in the past form a clade, which we will also call a *cluster*. The number of taxa in a randomly selected cluster will be a random variable. The *cluster size distribution* (CSD) is a function of a threshold TMRCA t, and describes the probability of having a size m cluster if a lineage (i.e. branch) at time t is selected uniformly at random from the set of all lineages at t and the size of the cluster descended from that branch is counted. A schematic of how clusters and the CSD are constructed given a tree and a threshold is shown in figure S5. In [6] we derived differential equations that describe the moments of the CSD.

Some of the properties of phylogenies that we seek to reproduce with the model developed below are:

Table 1. Epidemiological parameters.

Parameter	Symbol	Value
Transmission rate of early/acute	β_1	1 per 47 days
Transmission rate of chronic	β_2	1 per 1207 days
Mean duration of risk behavior	$1/\mu$	19.5 years
Mean duration of early/acute period	$1/\gamma_1$	180 days
Mean duration of chronic period	$1/\gamma_2$	10 years

doi:10.1371/journal.pcbi.1002552.t001

- The number of lineages as a function of time (NLFT), also known as the *ancestor function*.
- The fraction of sampled early/acute and chronic infections which are clustered given a threshold TMRCA.
- Within a given cluster there will a number of early/acute taxa and a number of chronic taxa. We will calculate the correlation coefficient between these counts across all clusters given a threshold TMRCA.
- The moments of the distribution of cluster sizes, including the mean, variance, and skew of cluster sizes.

Figure 1 shows a simple genealogy that could be generated by the HIV model. Four events can occur in this genealogy representing coalescence or the changing stage of a lineage. By quantifying the rate that these events occur using a coalescent model, we can calculate the clustering properties of these genealogies. These methods are described below and in detail in supporting Text S1.

The ancestor function is strictly decreasing in reverse time and converges to one (a single lineage) when the most recent common ancestor of the sample is reached. The initial value of the ancestor function (when the population is sampled) is equal to the sample size n. For the purposes of modeling phylogenetic properties of HIV, we will be interested in phylogenies such that the taxa are labeled with the state of the sampled individual (e.g. the individual will have early or late infection corresponding to the states in equation 1). In this case, we will have two ancestor functions, since a lineage may correspond to an infected individual with either early or late infection.

The ancestor functions derived from equations 1, and which are derived in the Text S1 are as follows:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}A_1 &= \gamma_1 I_1 \frac{A_2}{I_2} - \beta_1 S \frac{I_1}{N}\left(\frac{A_1}{I_1}\right)^2 \theta \\ &- \beta_2 S \frac{I_2}{N}\frac{A_1}{I_1}\theta\end{aligned} \qquad (2)$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}A_2 &= -\gamma_1 I_1 \frac{A_2}{I_2} \\ &+ \beta_2 S \frac{I_2}{N}\frac{A_1}{I_1}\frac{I_2 - A_2}{I_2}\theta.\end{aligned}$$

In these equations, A_1 is the number of lineages corresponding to early infections and A_2 is the number of lineages corresponding to late infections. These equations provide a deterministic approximation to the NLFT, which is $A(t) = A_1(t) + A_2(t)$. Each term in these equations accounts for loss or gain of lineages due to the concurrent processes of transmission (at rates $\beta_1 S \frac{I_1}{N}\theta$ and $\beta_2 S \frac{I_2}{N}\theta$) and transition between states (at rates $\gamma_1 I_1$). This approximation becomes exact in the limit of large sample and population size. Note that since the model is continuous in both time and state variables, the ancestor functions are not integers in contrast to most coalescent frameworks based on discrete mathematics.

Real epidemics in a finite population will have transmission trees such that the number of lineages at any time is a random variable. The mean-field model presented in equation 1 can be viewed as a description of the dynamics of a stochastic system in the limit of large population size. In this case, we can adapt the coalescent to make approximate descriptions of the stochastic properties of the transmission tree in large populations. The ancestor functions will reflect the approximation of the actual (random) number of lineages. Previous work has demonstrated that deterministic descriptions can be excellent approximations for the number of lineages over time [6,33]. In the following section, we compare our deterministic coalescent to stochastic simulations, confirming that it is a good approximation over a wide range of parameters.

Given a clustering threshold TMRCA t, the random variable $X_k(l;t)$ will be the number of stage-k taxa descended from a given lineage l that is extant at time t in the past. As before, $A_k(t)$ will be the number of type k lineages at the time t in the past. In our model, infected can be of two types (early/acute and chronic infected), so there are only two types: $k=1$ corresponds to earl/acute and $k=2$ corresponds to chronic. We will denote the set of all lineages of type k at time t in the past as $\mathcal{S}(k;t)$. Then we define the i and j'th moment of cluster sizes descended from a type k lineage to be

$$M_{i,j}(k;t) = \frac{1}{A_k(t)} \sum_{l \in \mathcal{S}(k;t)} X_1^i(l;t) X_2^j(l;t). \qquad (3)$$

Many summary statistics that are potentially informative about transmission dynamics can be derived from these moments. The moments are difficult to interpret, so in practice we use them to calculate summary statistics such as variance and skew of the CSD. Below, we examine 30 summary statistics derived from the first three moments and multiple clustering thresholds.

For example, the variance of cluster sizes counting only type 1 taxa descended from type k lineages is

$$Var(X_1;k) = M_{2,0}(k) - (M_{1,0}(k))^2. \qquad (4)$$

The total variance of cluster sizes counting only stage 1 taxa is found with the weighted average over lineage types:

$$Var(X_1) = \frac{A_1 Var(X_1;1) + A_2 Var(X_1;2)}{A_1 + A_2}. \qquad (5)$$

A similar set of equations can be developed for the cluster sizes aggregated over taxon types, that is, for $X_1 + X_2$. Detailed derivations are provided in Text S1 for differential equations that describe these moments as function of the threshold t.

Event-driven stochastic simulations were conducted to verify the suitability of the deterministic approximations for inference. Simulations implemented a variation on the Gillespie algorithm [34]. Populations consisted of $N = 5 \times 10^4$ agents, and were simulated for 15 or 30 years starting with one hundred initial infections. At the end of each simulation, a sample of either 20%

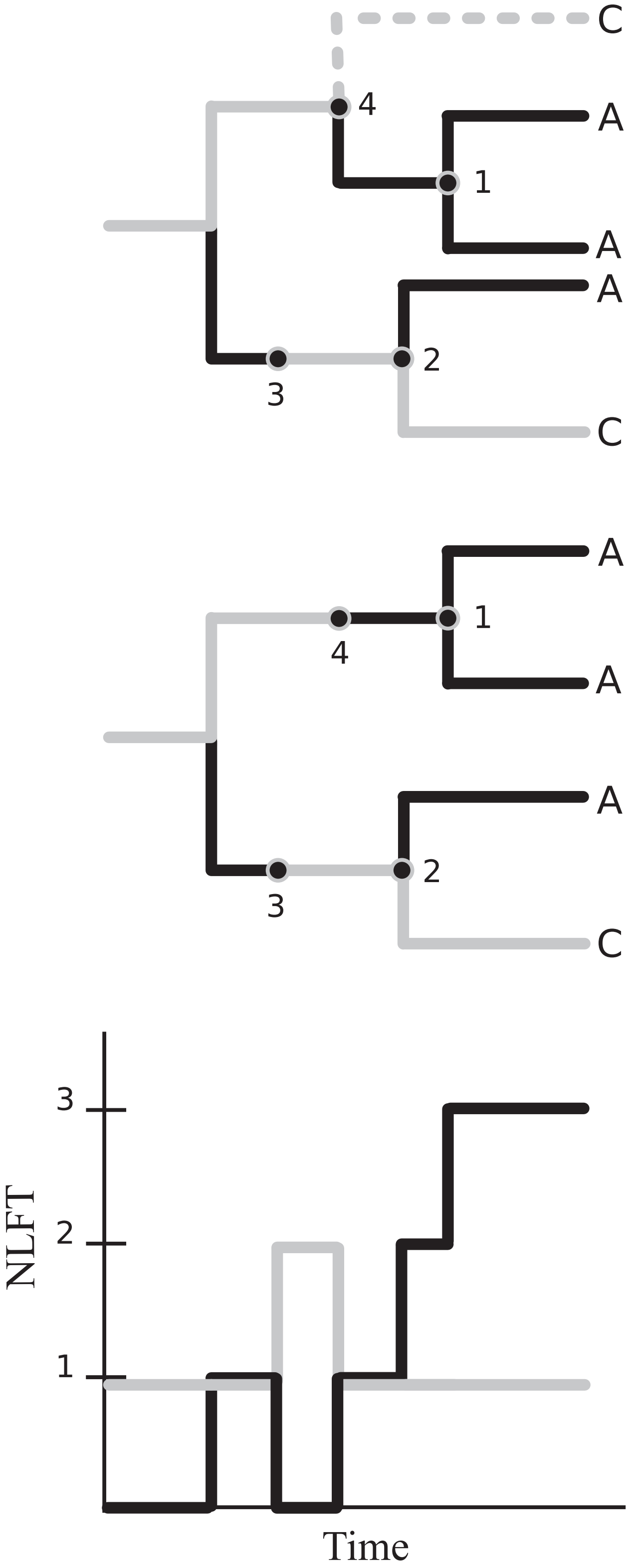


Figure 1. A simple gene genealogy that could be generated by the HIV model. Dark branches with taxa labeled *A* correspond to stage-1 (early/acute infected hosts). Light branches with taxa labeled *C* correspond to stage-2 (chronic infections). Event 1 represents the coalescence of two lineages corresponding to early/acute infection. Event 2 represents coalescence of an early and a late infection. Event 3 represents the stage transition of an early infection to a late infection. Event 4 represents the transmission by a late infection which is not ancestral to the sample. Top: Includes an unsampled lineage (dashed). Middle: The unsampled lineage has been pruned from the tree. The point where the lineage is pruned corresponds to event 4. Bottom: The number of lineages as a function of time (NLFT) which correspond to a host with early/acute infection (black) or chronic infection (grey).
doi:10.1371/journal.pcbi.1002552.g001

or 100% of prevalent infections was taken and used to reconstruct a transmission tree. Five hundred simulations were conducted for each sample fraction and sample time. Corresponding to each simulation, 10 transmission trees were generated based on a random sampling of using distinct clustering thresholds. The CSDs were then estimated from each tree and the moments of these distributions were compared to the moment equations (3–5).

We have further conducted an investigation into the potential of various summary statistics of the viral phylogeny for inference of underlying epidemiological parameters. Of particular interest is the fraction of transmissions that occur during early HIV infection. As indicated above, it is possible that phylogenetic clustering of early infections reflects elevated transmission during early/acute HIV infection, which we will define as the infectious period from zero to six months. The following simulation experiment was carried out to identify informative statistics:

1. Parameters β_l,β_h,N were sampled from a multivariate uniform distribution. 1800 replicates were sampled.
2. For each set of parameters, the HIV ODE model was integrated. The number of transmissions by early/acute and chronic cases was recorded. The number of stage transitions from acute to chronic was also recorded.
3. For each record of transmissions and stage transitions, a coalescent tree was simulated using the method described in [35].
4. For each coalescent tree, summary statistics were calculated and recorded. These statistics consisted of the following: The number of lineages as a function of time before the most recent sample; the correlation between between the number of early/acute and chronic infections with threshold TMRCA; the fraction of acute/recent taxa which remain unclustered (not clustered with any other taxa); the fraction of chronic taxa which remain unclustered; the mean number of taxa clustered with a early/acute infection; the mean number of taxa clustered with a chronic infection. Each of these statistics was calculated using 5 threshold TMRCA uniformly distributed between one year and 25 years before the most recent sample.

The coalescent tree was simulated such that the sample size matched that of the Detroit MSM phylogeny, and the heterochronous sampling of that phylogeny was reproduced in the coalescent tree. Furthermore, the number of early/acute versus chronic taxa sampled was determined using the BED test for recency of infection for each patient [36], and simulations were also made to match the numbers of early/acute and chronic taxa sampled. Virus from patients with early/acute infection accounted for 24% of the samples.

Summary statistics were centralized around the mean and rescaled by their standard deviation ($\frac{X-\mathrm{E}[X]}{\sigma(X)}$). The dependent variable of interest is the fraction of transmissions attributable to the acute stage at the beginning of the epidemic, which may be defined

$$\begin{aligned}\tau &= R_0^1/R_0 \\ &= \frac{\beta_1/\gamma_1}{\beta_1/\gamma_1+\beta_2/\gamma_2},\end{aligned} \tag{6}$$

where R_0^1 is the expected number of transmissions generated during early/acute infection at the beginning of the epidemic, and R_0 is the expected number of transmissions over the entire infectious period. Pearson correlation coefficients were calculated for each statistic and τ. To give a better indication which statistics would be useful for estimating the ratio of acute to chronic transmission rates, we conducted a partial least-squares (PLS) regression [37], which has been used by other investigators when estimating parameters by approximate Bayesian computation (ABC) methods [38]. Prediction error was assessed with 10-fold cross validation. We controlled for the sample fraction by including the prevalence of infection at the time of the most recent sample as a covariate.

Results

The mean CD4 cell count and FAS for clustered taxa is shown in figure 2. Consistent with our hypotheses, patients with higher CD4 count are more likely to yield phylogenetically clustered virus, and the mean CD4 count among clustered patients has an inverse relationship with the threshold TMRCA for clustering. Also consistent with our hypothesis, patients which yield virus with lower FAS (less diverse virus) are more likely to be phylogenetically clustered, and mean FAS has a positive relationship with the threshold TMRCA for clustering. Patients were strongly co-clustered within quantiles. Maximum assortativity values, which measures the similarity of co-clustered taxa were 13% for CD4 and 4.5% for FAS. The maximum assortativity also occurs at low threshold TMRCA for FAS and CD4 (1700 and 1467 days). Very little clustering is observed between the first and last quantiles.

In general, the deterministic model offers an excellent approximation to the stochastic system. All trajectories pass through or close to the median of simulation predictions. Figure 3 illustrates the prevalence of early/acute and chronic infections from a typical simulation of the HIV model and the corresponding deterministic approximations. This correspondence occurs despite large fluctuations in prevalence when the number of infections is small. In [6] it was shown that the correspondence between the stochastic and deterministic systems can be very good even if the epidemic is started from a single infection and the coalescent is fit to the resulting transmission tree.

In figure 3, late infections outnumber early infections by approximately 20 to 1. As a consequence, NLFT for late infections are more stable due to larger sample sizes, and the NLFT are more noisy for the sample of early infections. The prevalence of infection plateaus prior to the 15 year sample time, so there is not much difference in the phylogenetic features observed at 15 and 30 year sampling times.

Many summary statistics calculated from an HIV gene genealogy can be informative about the fraction of transmissions attributable to early/acute infection, τ (equation 6). Figure 4 shows the value of four statistics as τ is varied. The dependancy of these summary statistics on the sample fraction is also shown in figure S4. $\rho(X_1,X_2)$ (upper left) is the Pearson correlation coefficient between the number of early/acute taxa and chronic taxa in a cluster and is most sensitive to τ. Also shown are the mean cluster size, the number of extant lineages at the threshold TMRCA, and the fraction of taxa in a phylogenetic cluster. As the fraction of transmissions from the early/acute stage is varied, transmission rates β_1 and β_2 are adjusted so that R_0 remains constant. The smallest value of τ shown in figure 4 corresponds to the point where $\beta_1=\beta_2$, such that there is no excess transmission in the early/acute stage. The most recent sample is assumed to be at 35

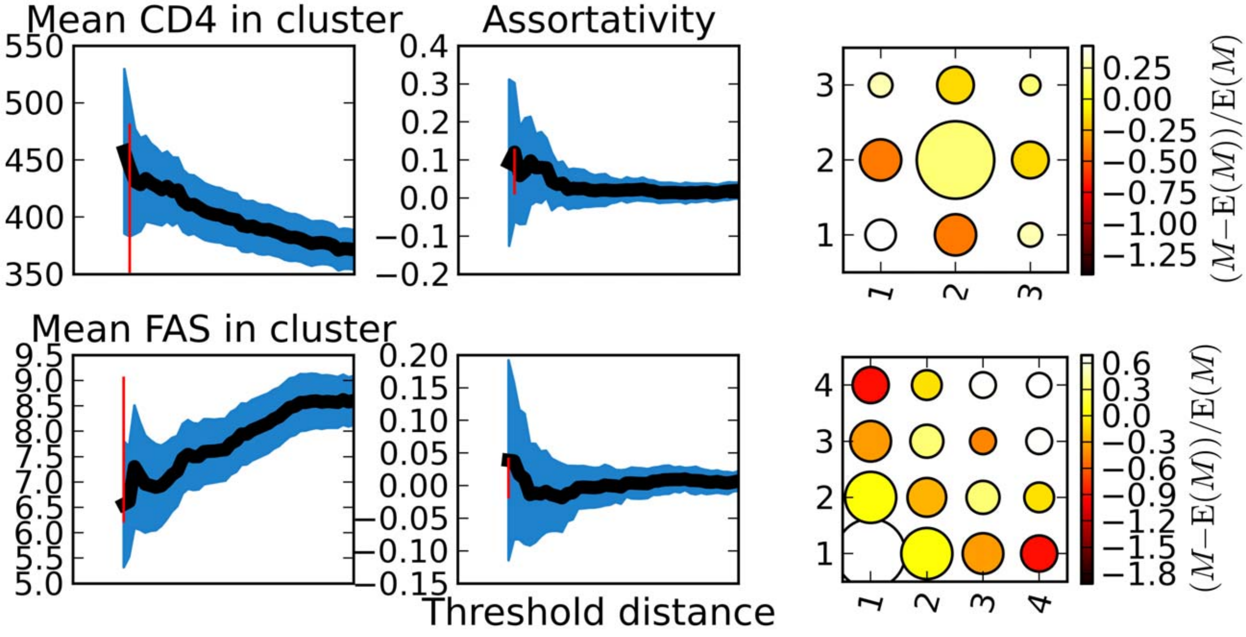


Figure 2. Excess clustering and excess co-clustering of virus from patients with early/acute infections. Left: The mean CD4 cell count (top) and frequency of ambiguous sites (bottom) versus the threshold TMRCA used to form clusters. Middle: The assortativity coefficient, a measure of similarity of co-clustered taxa, versus the treshold TMRCA used to form clusters. Assortativity of CD4 is at top, and frequency of ambiguous sites is bottom. Right: The size of each matrix element is proportional to number of co-clusterings between taxa categorized by CD4 (top, $x_1<200<x_2<800<x_3$) or quartile of frequency of ambiguous sites (bottom). The color represents the extent to which the count of co-clusterings exceeds the expectation if clusters were forming at random. The color scale (far right) shows strong assortativity within quartiles. The vertical red bar represents the threshold which was used to create clusters and the matrix derived from the set of clusters. This threshold corresponds to the maximum of the assortativity coefficient for the derived matrix.
doi:10.1371/journal.pcbi.1002552.g002

years following the initial infection. Epidemic prevalence after 35 years is approximately constant. The threshold TMRCA was five years before the most recent sample. Sample size and distribution of samples over time was matched to the Detroit MSM phylogeny. Furthermore, the number of early/acute versus chronic taxa sampled was made to match the Detroit data by use of the BED test [36] for determining recency of infection.

The fraction of taxa which are phylogenetically clustered also varies with τ (figure 4, upper left). The fraction of early/acute taxa clustered is more sensitive to τ than the fraction from chronic taxa. Early/acute taxa are always clustered at a greater rate than chronic taxa, even when $\beta_1 = \beta_2$ corresponding to the minimum value of τ. This is because virus from early/acute patients was recently transmitted, making it much more likely that the lineage will coalesce in the recent past regardless of the source of the infection.

Using the mathematical model, we explored many parameters including the threshold TMRCA for clustering, the sample fraction, and the time relative to the beginning of the epidemic at which sampling occurs. Figures S1, S2, S3 demonstrate that the deterministic model is capable of reproducing many phylogenetic signatures that have been associated with HIV epidemics in MSM. For example, figure S5 shows the fraction of the sample (both early and late infections) which remain unclustered with any other sample unit. When the threshold TMRCA is zero (corresponding to the far right of the time axis), the entire sample remains unclustered. As the threshold TMRCA increases (moving leftwards on the time axis), more sample units become clustered and the fraction of taxa remaining unclustered decreases.

The time of sampling makes little absolute difference to the qualitative nature of the tree statistics if sampling occurs after the peak epidemic prevalence (around 15 years). However the sample fraction (the fraction of prevalent infections sampled) has a large effect on all tree statistics. When the sample fraction is large, the fraction remaining unclustered drops much more precipitously than when it is small as the threshold TMRCA increases. This occurs because each transmission can cause a sample unit to become clustered; a large sample size implies that transmissions will have a greater probability of resulting in an observable coalescent event (e.g. it results in a larger ratio A_i/I_i).

Early infections become clustered at a much greater rate than late infections. This corresponds to the excess clustering of early/acute infections observed in many phylogenies. By virtue of being infected in the recent past, an acute infection inevitably has a very recent common ancestor with another infection who transmitted to that individual. Mathematically, this is reflected in transmission terms of the form $\beta_1 S (I_1/N)(A_1/I_1)^2$ which appear in the ancestor function for early, but not late infections.

When the sample fraction is non-negligible, the fraction of the sample in a cluster levels off for intermediate thresholds. Similar phenomena were noted by Lewis et al. [8] and Hughes et al. [14] who observed that the fraction of the sample in a cluster did not change substantially beyond a small threshold, though these studies probably had high sample fractions. The plateau is due to the bimodality of coalescence times induced by early infection dynamics. Many coalesce events occurs at thresholds close to the sampling time, which corresponds to lineages of early infection coalescing.A larger group of coalescence times occurs close to the beginning of the epidemic when the effective population size is small. We hypothesize that the amount of excess clustering of early infections can be informative for estimating the sample fraction when it is not known.

Figure S2 shows the Pearson correlation coefficient for the number of co-clustered early and chronic infections as a function of the clustering threshold ($\rho(X_1(l), X_2(l))$). Given that a sample unit is in a cluster, under certain circumstances, it is much more likely to be clustered with another unit of the same type. This is reflected by large negative correlation coefficients for the number of co-clustered early and late infections for small threshold TMRCA. But negative correlation between the number of early and late infections is only observed for small sample fractions and small threshold TMRCA. The region of negative correlation appears very briefly for a 100% sample fraction; the region is much longer for small samples. This implies that if a patient with early infection is clustered, it is much more likely to be clustered with another early infection than expected by chance alone.

The skewness of the CSD shows a similar trend (figure S3). The skewness is always positive (to the right) and rapidly decreases as the threshold TMRCA is increased reflecting greater probability mass in the tail of the distribution. Skew is greatest for small threshold TMRCA, when most clusters are of size 1. The distribution remains positively skewed, though it quickly levels off for intermediate threshold TMRCA. The mathematical model shows that all moments of the CSD are finite and diverge to infinity in the limit of large sample size and threshold TMRCA.

A practical consequence of having an intermediate to large sample fraction is that chains of acute-stage transmission will account for many of the clusters observed at low thresholds. If a taxon is clustered with an early infection, then it is *more* likely that the unit will be clustered with additional early infections since such cases are highly infectious and have likely transmitted in the recent past. This provides a justification for the theory expounded in Lewis et al. [8] that high clustering of cases with recent MRCA's indicates episodic transmission; chains of transmission by early infections are interrupted by occasional long intervals until a transmission by late stage infections.

Corroborating figure 4 which shows that many statistics are correlated with τ, the PLS regression did not single out any particular group of statistics as being informative of early/acute stage transmission rates. The first component distinguishes between statistics that describe co-clustering (correlation of the number of acute and chronic taxa in a cluster) and statistics that describe excess clustering (e.g. the fraction of early/acute taxa that are not clustered with any other taxa). Four principal components were required to explain 42% of the variance of the transmission fraction with additional components only explaining an additional 2%. All statistics were well represented in the model with four components.

Discussion

We have used coalescent models to characterize the phylogenetic patterns of a virus which produces an early stage of intensified transmission followed by a long period of low infectiousness. These patterns have been observed in multiple phylogenies of HIV-1 from MSM and IDU, and our model suggests that these should be general features for epidemics which feature early and intense transmission. These patterns are not necessarily a consequence of complex sexual network structure [14]. Complex transmission dynamics driven by sexual networks are undoubtedly taking place, but detecting the phylogenetic signature of sexual network structure will require carefully-chosen summary statistics [15]. We have characterized phylogenies using the cluster size distribution (CSD) which is similar to commonly used clustering methods based on strong support for monophyly but is nevertheless tractable for mathematical modeling in a dynamical systems framework. Moments of the CSD reflect a wide range of tree topologies, such as the distribution of branch lengths

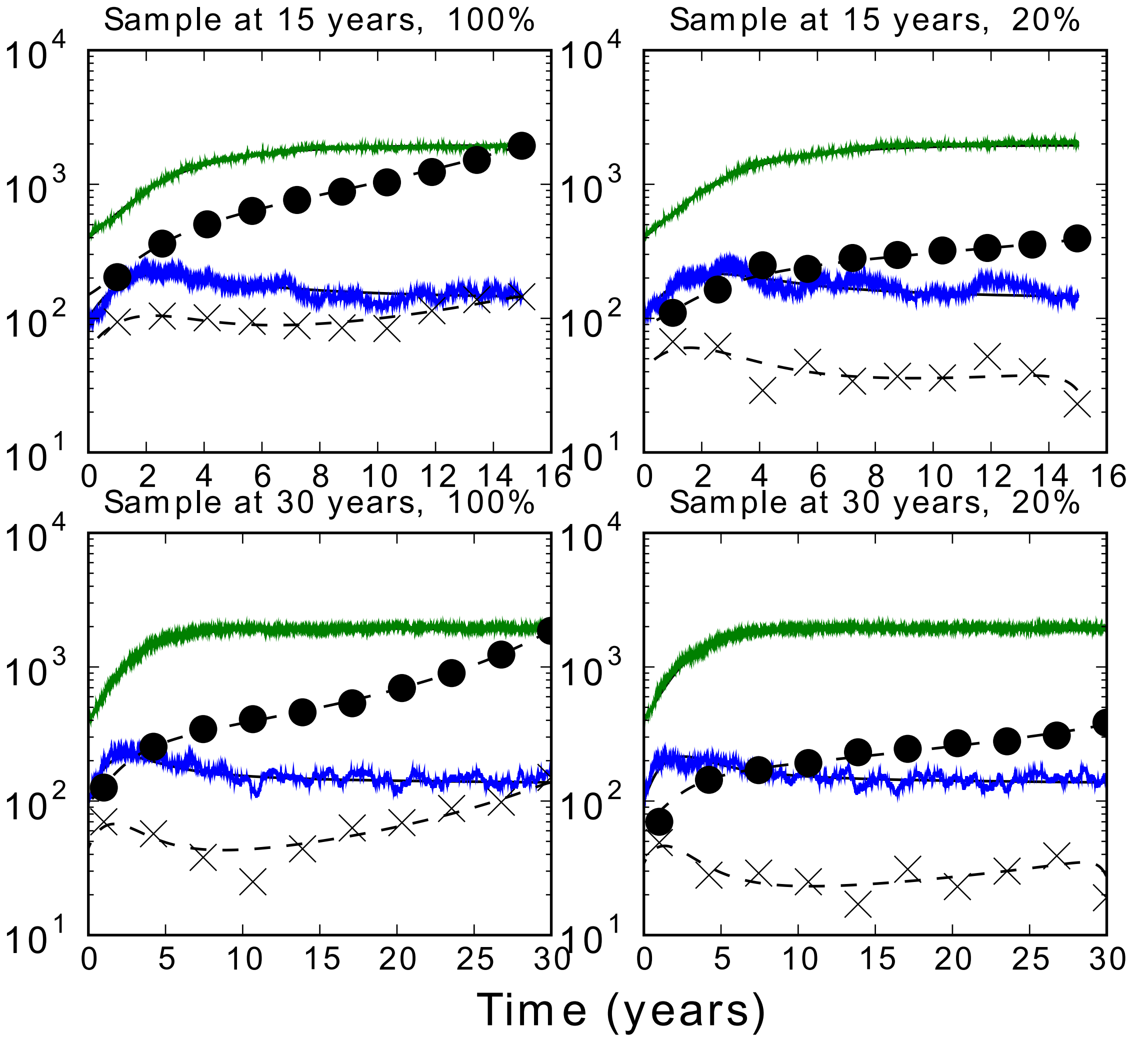


Figure 3. Two simulated epidemics and the deterministic approximations for the prevalent number of early and late infections and the ancestor functions (the number of lineages over time). The x-axis gives the time since the beginning of the epidemic, or equivalently, the threshold TMRCA used to calculate the number of lineages over time. Green describes the simulated number of late infections. Blue describes the simulated number of early infections. Dots show the simulated ancestor function for the number of lineages that correspond to late infections. And x's show the simulated ancestor function for lineages in early infection. Dashed lines show the prediction of the deterministic coalescent. The top row shows results for a sample taken at 15 years following the initial infections, and the bottom shows results for a sample at 30 years. The right column shows results for a sample fractions of 20%, and the left column for a census of prevalent infections(100%).
doi:10.1371/journal.pcbi.1002552.g003

and tree balance, and are potentially informative of a wide range population genetic processes. For example, a highly unbalanced tree would have produce very skewed CSD, and a very star-like tree would have a CSD that is insensitive to changes in the clustering threshold.

While there has been much discussion of how clustering of acute infections is caused by the intensity of transmission during the acute stage, the amount of excess clustering that will be observed is also very sensitive to the sample fraction. And even if transmission rates in the early/acute stage are equal to those in the late/chronic stage, we would still observe excess clustering of early/acute provided the sample fraction was large enough. This is a simple consequence of early/acute infections being connected by short branch lengths to the individual who transmitted infection. An advantage of the coalescent framework used in this investigation is that it is accurate even with large sample fractions [35].

Some of the statistics which are most informative of the underlying epidemiological processes are those based on co-clustering of labeled taxa, such as the correlation between the number of early and late infections in a cluster. Such statistics tend to be the most responsive to variation of the intensity of transmission during early infection, and are therefore good candidates for future estimation of the fraction of transmissions that occur during the first few months of infection with HIV. Knowing the frequency of early transmission is essential to prevention efforts, since these transmissions are the most difficult to prevent. Individuals with early and acute infection are usually not aware of the infection, and are therefore not susceptible to many interventions. Modeling to evaluate strategies such *seek, test, and treat* (STT) [39,40] and *pre-exposure prophylaxis*(PrEP) [41] will require good estimates for the frequency of early-stage transmission in diverse populations, and phylogenetic data promise to refine these estimates.

Future work could focus on finding ways to use statistics derived from the CSD for estimation of epidemiological parameters within an approximate Bayesian framework [38,42,43]. Alternatively,

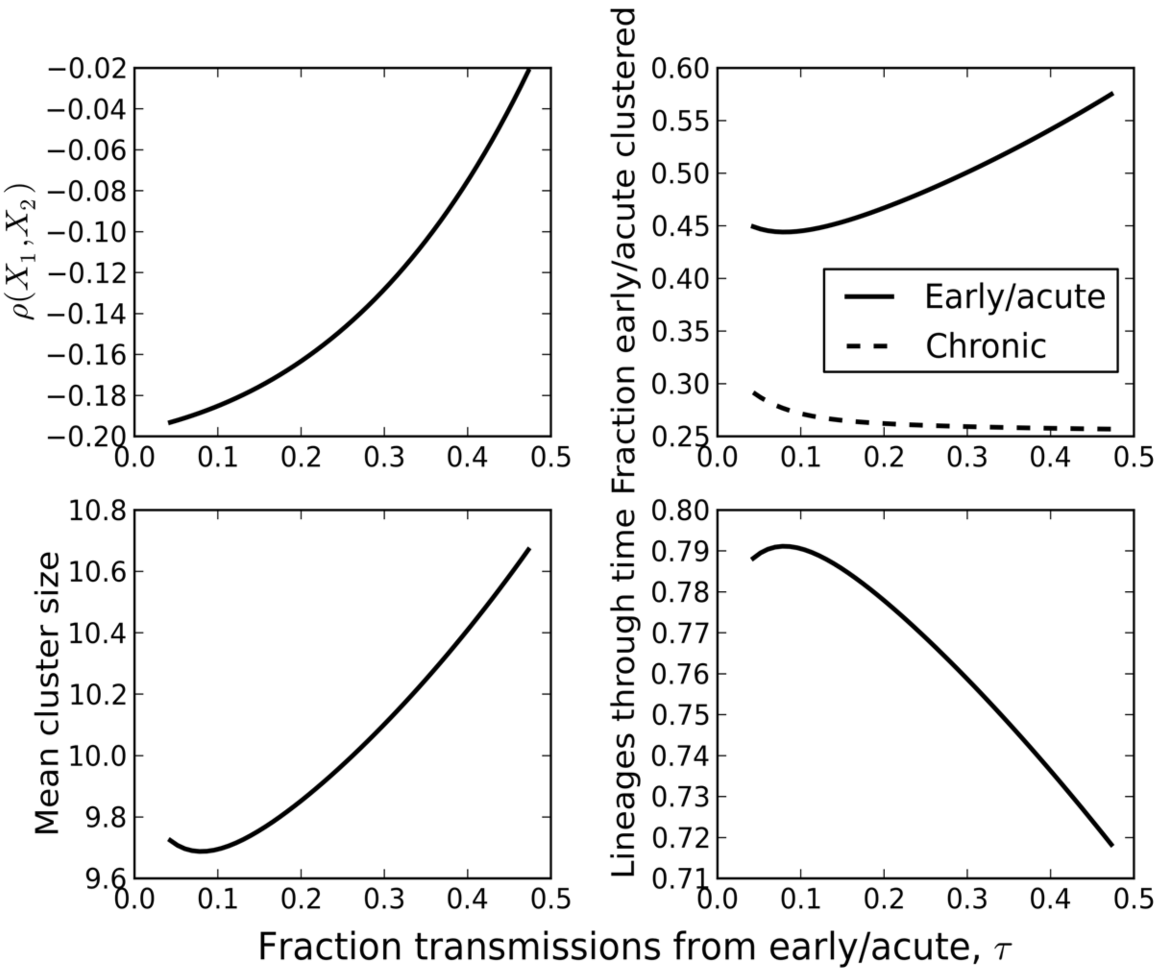


Figure 4. Summary statistics from HIV gene genealogies versus the fraction of transmissions attributable to early/acute infection. The threshold TMRCA was five years before the most recent sample. Sample size and distribution of samples over time was matched to the Detroit MSM phylogeny.
doi:10.1371/journal.pcbi.1002552.g004

advances [35] in coalescent theory may make it possible to calculate the likelihood of a gene genealogy conditional on a complex demographic history, such as those generated by the HIV model discussed here. Current techniques are limited in the amount of phylogenetic data that can be used for inference of demographic and epidemiological parameters. Estimation of the intensity of early stage transmission will likely require co-clustering statistics similar to the moments derived from the CSD. In cases where the simple compartmental models fail to reproduce phylogenetic patterns, a more complex transmission system model and its corresponding coalescent should be investigated which might involve sexual networks or geographical [44] and risk heterogeneity. We further conclude that care must be taken in using phylogenetic clusters for epidemiological inference. Mechanisms that generates clustering are often complex and counter-intuitive. We recommend that investigators shift from individual-based inference using small clusters to model-based inference using population-based surveys of sequence diversity.

Supporting Information

Figure S1 Two simulated epidemics and the deterministic approximations for the fraction of the sample which remains un-clustered as a function of the threshold TMRCA. The fraction un-clustered is shown for sample units classified as early infections (solid lines) as well as sample units that are late infections (dashed). The x-axis gives the clustering threshold in units of days since the start of the epidemic. All variables are illustrated for a sample at 30 years following the initial infections and at two possible sample fractions (100% or 20%).
(EPS)

Figure S2 Simulated epidemics and the deterministic approximations for the Pearson correlation coefficient between the number of co-clustered early and late infections. Variables are shown as a function of the threshold TMRCA in units of days since the beginning of the epidemic. All of these variables are illustrated for a sample at 30 years following the initial infections and at two possible sample fractions (100% or 20%).
(EPS)

Figure S3 Two simulated epidemics and the deterministic approximations for the skewness of the cluster size distribution (third central moment divided by the standard deviation cubed). Variables are shown as a function of the threshold TMRCA in units of days since the beginning of

the epidemic. All variables are illustrated for a sample at 30 years following the initial infections and at two possible sample fractions (100% or 20%).
(EPS)

Figure S4 Summary statistics from HIV gene genealogies versus the fraction of infections sampled after 35 years. The threshold TMRCA was five years before the most recent sample. Sampling was homochronous.
(EPS)

Figure S5 Construction of the cluster size distribution (CSD). Given a tree and a threshold time to most recent common ancestor, represented by red, green, and blue lines, the set of taxa at the base of the tree are classified into disjoint sets or *clusters*. The distribution of cluster sizes for each threshold is shown at right.
(EPS)

Text S1 Detailed derivations and simulation methods.
(PDF)

Acknowledgments

The authors thank Eve Mokotoff and Mary-Grace Brandt and colleagues at the Michigan Department of Community Health for assisting with access to the HIV drug-resistance database.

Author Contributions

Conceived and designed the experiments: EMV SDWF. Performed the experiments: EMV. Analyzed the data: EMV SDWF. Contributed reagents/materials/analysis tools: EMV SDWF. Wrote the paper: EMV JSK MJW ALB SDWF.

References

1. Bruisten S, Schouls L (2010) Molecular typing and clustering analysis as a tool for epidemiology of infectious diseases. Modern Infectious Disease Epidemiology: Concepts, Methods, Mathematical Models, and Public Health : 117.
2. Pybus O, Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet 10: 540–50.
3. Grenfell B, Pybus O, Gog J, Wood J, Daly J, et al. (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303: 327.
4. Zhu T, Mo H, Wang N, Nam D, Cao Y, et al. (1993) Genotypic and phenotypic characterization of HIV-1 patients with primary infection. Science 261: 1179.
5. Li H, Bar K, Wang S, Decker J, Chen Y, et al. (2010) High multiplicity infection by HIV-1 in men who have sex with men. PLoS Pathog 6: e1000890.
6. Volz E, Pond S, Ward M, Leigh Brown A, Frost S (2009) Phylodynamics of Infectious Disease Epidemics. Genetics 183: 1421–30.
7. Wilson DJ, Falush D, McVean G (2005) Germs, genomes and genealogies. Trends Ecol Evol 20: 39–45.
8. Lewis F, Hughes G, Rambaut A, Pozniak A, Leigh Brown A (2008) Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med 5: 392–402.
9. Brenner B, Roger M, Routy J, Moisi D, Ntemgwa M, et al. (2007) High rates of forward transmission events after acute/early HIV-1 infection. J Infect Dis 195: 951–959.
10. Pao D, Fisher M, Hué S, Dean G, Murphy G, et al. (2005) Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. AIDS 19: 85.
11. Brenner B, Roger M, Moisi D, Oliveira M, Hardy I, et al. (2008) Transmission networks of drug resistance acquired in primary/early stage HIV infection. AIDS 22: 2509.
12. Yerly S, Junier T, Gayet-Ageron A, Amari E, von Wyl V, et al. (2009) The impact of transmission clusters on primary drug resistance in newly diagnosed HIV-1 infection. AIDS 23: 1415.
13. Cuevas M, Muñoz-Nieto M, Thomson M, Delgado E, Iribarren J, et al. (2009) HIV-1 transmission cluster with T215D revertant mutation among newly diagnosed patients from the Basque Country, Spain. J Acquir Immune Defic Syndr 51: 99.
14. Hughes G, Fearnhill E, Dunn D, Lycett S, Rambaut A, et al. (2009) Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathog 5: e1000590.
15. Leigh Brown A, Lycett S, Weinert L, Hughes G, Fearnhill E, et al. (2011) Transmission network parameters estimated from HIV sequences for a nationwide epidemic. J Infect Dis 204: 1463–9
16. Liljeros F, Edling C, Amaral L, Stanley H, Åberg Y (2001) The web of human sexual contacts. Nature 411: 907–908.
17. Pilcher C, Tien H, Eron Jr J, Vernazza P, Leu S, et al. (2004) Brief but efficient: acute HIV infection and the sexual transmission of HIV. J Infect Dis 189: 1785–1792.
18. Koopman J, Jacquez J, Welch G, Simon C, Foxman B, et al. (1997) The role of early HIV infection in the spread of HIV through populations. J Acquir Immune Defic Syndr 14: 249.
19. Bezemer D, van Sighem A, Lukashov V, van der Hoek L, Back N, et al. (2010) Transmission networks of HIV-1 among men having sex with men in the Netherlands. AIDS 24: 271.
20. Kim J, Riolo R, Koopman J (2010) HIV transmission by stage of infection and pattern of sexual partnerships. Epidemiology 21: 676.
21. Kuiken C, Leitner T, Foley B, Hahn B, Marx P, et al. (2009) HIV sequence compendium 2009. Los Alamos, New Mexico: Los Alamos National Laboratory, Theoretical Biology and Biophysics.
22. Rose P, Korber B (2000) Detecting hypermutations in viral sequences with an emphasis on g-a hypermutation. Bioinformatics 16: 400–401.
23. Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
24. Bennett D, Camacho R, Otelea D, Kuritzkes D, Fleury H, et al. (2009) Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. PLoS One 4: e4724.
25. Drummond A, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7: 214.
26. Taffé P, May M (2008) A joint back calculation model for the imputation of the date of HIV infection in a prevalent cohort. Stat Med 27: 4835–4853.
27. Kouyos R, von Wyl V, Yerly S, Böni J, Rieder P, et al. (2011) Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. Clin Infect Dis 52: 532.
28. Shankarappa R, Margolick J, Gange S, Rodrigo A, Upchurch D, et al. (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 73: 10489.
29. Newman M (2003) Mixing patterns in networks. Phys Rev E 67: 026126.
30. Frost S, Volz E (2010) Viral phylodynamics and the search for an effective number of infections. Philos Trans R Soc Lond B Biol Sci 365: 1879.
31. Hein J, Schierup M, Wiuf C (2005) Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford University Press, USA.
32. Wakeley J (2005) Coalescent theory: an introduction. Roberts Greenwood Village, Colorado.
33. Maruvka Y, Shnerb N, Bar-Yam Y, Wakeley J (2011) Recovering population parameters from a single gene genealogy: An unbiased estimator of the growth rate. Mol Biol Evol 28: 1617.
34. Gillespie D (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81: 2340–2361.
35. Volz E (2012) Complex population dynamics and the coalescent under neutrality. Genetics 190: 187–201.
36. Prejean J, Song R, Hernandez A, Ziebell R, Green T, et al. (2011) Estimated HIV incidence in the united states, 2006–2009. PloS One 6: e17502.
37. Mevik B, Wehrens R (2007) The pls package: Principal component and partial least squares regression in R. J Stat Softw 18: 1–24.
38. Wegmann D, Leuenberger C, Excoffier L (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182: 1207.
39. Granich R, Gilks C, Dye C, De Cock K, Williams B (2009) Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. The Lancet 373: 48–57.
40. Wood E, Kerr T, Marshall B, Li K, Zhang R, et al. (2009) Longitudinal community plasma HIV-1 RNA concentrations and incidence of HIV-1 among injecting drug users: prospective cohort study. BMJ 338: b1649.
41. Grant R (2010) Antiretroviral agents used by HIV-uninfected persons for prevention: pre-and postexposure prophylaxis. Clin Infect Dis 50: S96.
42. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf M (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interface 6: 187.
43. Beaumont M, Cornuet J, Marin J, Robert C (2009) Adaptive approximate Bayesian computation. Biometrika. DOI: 10.1093/biomet/asp052
44. Lemey P, Rambaut A, Drummond A, Suchard M (2009) Bayesian phylogeography finds its roots. PLoS Comput Biol 5: e1000520.