Phylodynamics of Influenza A viruses

by

Leila Rahnama

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
for a Master (M.Sc.) degree in
Bioinformatics

The Ottawa-Carleton Institute of Biology &
Department of Biology
Faculty of Science
University of Ottawa

Abstract

Human populations are constantly exposed to emerging pathogens such as influenza A viruses that result from cross-species transmissions. Generally these sporadic events are evolutionary dead-ends, but occasionally, viruses establish themselves in a new host that offers a novel genomic context to which the virus must adjust to avoid attenuation. However, the dynamics of this process are unknown. Here we present a novel method to characterize the time it takes to G+C composition at third codon positions (GC3 content) of influenza viruses to adjust to that of a new host. We compare the inferred dynamics in two subtypes, H1N1 and H3N2, based on complete genomes of viruses circulating in humans, swine and birds between 1900-2009. Our results suggest that both subtypes have the same fast-adjusting genes, which are not necessarily those with the highest absolute rates of evolution, but those with the most relaxed selective pressures. Our analyses reveal that NA and NS2 genes adjust the fastest to a new host and that selective pressures of H3N2 viruses are relaxed faster than for H1N1. The asymmetric nature of these processes suggests that viruses with the greatest adjustment potential to humans are coming from both birds and swine for H3N2, but only from birds for H1N1.

Acknowledgements

Many thanks to all whose supports were critical to the development and completion of this thesis. I am sincerely grateful to my supervisor, Dr. Stéphane Aris-Brosou, for giving me the opportunity to pursue my degree under his supervision. I was extraordinarily fortunate to work with him as a prominent expert in the field of bioinformatics. Stéphanes' insights, experiences and philosophy of science have provided me with new perspectives on practicing science and performing research.

Special thanks go to my great committee members Dr. Xuhua Xia and Dr. Ashkan Golshani for all their challenging questions, recommendations and intellectual supports.

Furthermore, I am specifically grateful for my exceptional siblings, for my beloved mother, Nosrat Yekkallam Maharlouei and my father, Behrooz Rahnama. I thank them for their unconditional love, excellent ideologies, and unlimited emotional and financial support, which enabled me to achieve my goals throughout my education. I dedicate this thesis to all of them.

List of Abbreviations

WHO	World Health Organization
NIH	National Institutes of Health
N	Non-Synonymous
S	Synonymous
Gal	Galactose
Sia	Sialic acid
vRNA	Viral RNA
RNA	Ribonucleic acid
RNP	Ribonucleoprotein (known also as the RNA-dependent RNA polymerase complex)
PB2	Basic polymerase 2
PB1	Basic polymerase 1
PA	Acidic polymerase
HA	Hemagglutinin
NP	Nucleocapsid protein
NA	Neuraminidase
M1	Matrix protein1
M2	Matrix protein2
NS1	Non-structural protein1
NS2	Non-structural protein2
NEP	Nuclear export protein
IFN	Interferon
cRNA	complementary RNA
RdRp	RNA dependent RNA polymerase

NJ	Neighbor-joining
ML	Maximum likelihood
BEAST	Bayesian Evolutionary Analysis by Sampling Trees

Contents

1 Introduction **1**

1.1 Epidemiology of influenza 1

1.1.1 Pandemics and epidemics of influenza 1

1.1.2 Global pandemics of influenza history 2

1.1.3 Public health importance of influenza 3

1.1.4 Influenza transmission and required receptors 4

1.2 Molecular virology of influenza 6

1.2.1 Classification 7

1.2.2 Hosts 7

1.2.3 Genome: structure and functions 8

1.2.4 Life cycle 15

1.3 Molecular evolution of influenza 16

1.3.1 Origin of influenza viruses 16

1.3.2 Genomics in studying the influenza virus evolution 18

1.3.3 Phylogenetic approaches to study evolution of influenza viruses . 19

1.3.4 The mechanisms of influenza virus evolution 20

1.3.5 Major evolutionary changes of influenza viruses 21

1.3.6 Genetic changes 23

1.4 Objectives of the thesis 25

2 Methods **27**

2.1 Data collection and alignment 27

2.2 Sequence clustering 28

2.3 Phylogenetic analyses 29

2.4 Timing GC3 adjustment after a host change 30

2.5 Detection of selection 32

3 Results **34**

3.1 Sequence clustering 34

3.2 H1N1 and H3N2 subtypes evolve with extensive reassortment 35

3.3 Some genes evolve faster in H3N2 than in H1N1 36

3.4 Estimation of GC3 adjustment times 37

3.5 GC3 adjustment is faster in H3N2 than in H1N1 40

3.6 GC3 adjustment reflects relaxed selective pressures 44

4 Conclusions **47**

4.1 Principal findings 47

4.2 Future directions 50

4.3 Concluding statements 51

Bibliography **53**

Appendices **70**

List of Tables

2.1 **Genomic data used in this study.** *sl*: sequence length (nucleotides); ns_{total}: number of sequences before clustering; ns_{clust}: number of sequences after clustering with `DOTUR`. 28

2.2 **Best-fit substitution models of evolution.** Models were selected based on the Akaike Information Criterion (AIC). 30

2.3 **Significance of the trend observed in GC3 clusters.** n_{MSS} is the number of clusters estimated by median split silhouettes. P-values are derived from robust linear regressions. Significant results ($\alpha < 0.01$) are in bold. Cluster identifiers are arbitrary. 31

2.4 **Branch-site test of positive selection during adjustment to a new host**. Hyp: hypothesis; ℓ: log-likelihood; *np*: number of parameters; P: P-value; $\hat{\omega}$: estimated ω value; $\hat{p}_{\omega}$: proportion of sites (with $\hat{\omega} < 1$ under H_0, or $\hat{\omega} > 1$ under H_1); na: not applicable; long dash (—): data not available. 33

3.1 **Tests of the molecular clock assumption**. ℓ_{noclock}: log-likelihood without the clock assumption; ℓ_{clock}: log-likelihood under the clock assumption; *ns*: number of sequences; X^2: test statistic (twice the log-likelihood difference); df: degree of freedom; P: P-value. 38

List of Figures

1.1 **Schematic representation of influenza A pandemics.** The past pandemics of influenza A viruses (in 1918, 1957, 1968, 1977 and 2009) are the result of cross-species transmission caused by influenza A viruses of different subtypes (H1N1, H2N2 and H3N2) through different evolutionary processes. The responsible evolutionary mechanisms for each pandemic is shown under each virus. It is likely that next pandemic would probably be a H5N1 or H7N9. Common names of the past pandemics and the virus origins are shown in parentheses at the top and bottom of each virus, respectively. [1]. 3

1.2 **Schematic figure of an influenza A virion.** Three proteins, HA, NA, and the M2 transmembrane proton channel, are anchored in the viral membrane, which is composed of a lipid bilayer. The large, external domains of HA and NA are the major targets for neutralizing antibodies of the host immune response. The M1 matrix protein is located below the membrane with its counterpart M42 (in some strains). Inside the virion, the eight RNA segments are packaged in a complex with nucleoprotein (NP) and the RNA polymerase complex [2]. The RNA polymerase complex also includes PB1-F2, PB1-N40 and PA-X protein segments in some strains. 9

1.3 **Three-dimensional structure of the ribonucleoprotein (RNP) complex.** The schematic figure shows the three-dimensional model for: (A) the viral RNP complex including PB2, PB1 and PA. The location of specific domains in the PB1 (green), PB2 (red) and PA (violet) subunits are indicated. (B) The front-view of the polymerases is presented with the docking of the PA(C)-PB1(N) dimer and the N-terminal PB1 peptide is indicated with an arrow and highlighted in green [3]. 10

3.1 **Posterior mean rates of evolution of H3N2 *vs.* H1N1 viruses.** Results are shown on a $\log_e$-$\log_e$ scale (in substitutions/site/year). The gray line represents the first bisector (line of equation $y = x$), while the red line represents the linear fit to the data. Bars: limits of the 95% Highest Posterior Densities. 37

3.2 **Estimation of adjustment times.** Schematic representation of the method developed to estimate adjustment times. A host-switch event occurred along the red branch, and a GC3 cluster change occurred along the blue branch. Time t flows from the past to the present (bottom axis), and divergence times are estimated for nodes (see vertical broken lines). The two durations of interest are $\max_t$ and $\min_t$. See text for details. . . 40

3.3 **GC3 adjustment times of H3N2 *vs.* H1N1 viruses.** Results are shown on a $\log_e$-$\log_e$ scale (in years). The gray line represents the first bisector (line of equation $y = x$), while the red line represents the linear fit to the data. Bars: SEMs (95% Highest Posterior Densities, not shown, tend to be larger – see Fig. S27). $\star$: the HA value for H3N2 was tentatively derived using branches around the root node. 42

3.4 **Factor effects in the linear model (ANOVA) that was fitted to adjustment times (in years).** The directions of host change are avian-to-human (A-H), avian-to-swine (A-S), human-to-avian (H-A), human-to-swine (H-S), swine-to-avian (S-A) and swine-to-human (S-H). Adjustment times are in years. See text for details. 45

S1 **Timed tree for subtype H1N1 gene PB2.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray. 71

S2 **Timed tree for subtype H1N1 gene PB1.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray. 72

S3 **Timed tree for subtype H1N1 gene PA.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray. 73

S4 **Timed tree for subtype H1N1 gene HA.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray. 74

S5 **Timed tree for subtype H1N1 gene NP.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray. 75

S6 **Timed tree for subtype H1N1 gene NA.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray. 76

S7 **Timed tree for subtype H1N1 gene M2.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray. 77

S8 **Timed tree for subtype H1N1 gene M1.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray. 78

S9 **Timed tree for subtype H1N1 gene NS2.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray. 79

S10 **Timed tree for subtype H1N1 gene NS1.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray. 80

S11 **Timed tree for subtype H3N2 gene PB2.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. 81

S12 **Timed tree for subtype H3N2 gene PB1.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. 82

S13 **Timed tree for subtype H3N2 gene PA.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. 83

S14 **Timed tree for subtype H3N2 gene HA.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. 84

S15 **Timed tree for subtype H3N2 gene NP.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. 85

S16 **Timed tree for subtype H3N2 gene NA.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. 86

S17 **Timed tree for subtype H3N2 gene M2.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. 87

S18 **Timed tree for subtype H3N2 gene M1.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. 88

S19 **Timed tree for subtype H3N2 gene NS2.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. 89

S20 **Timed tree for subtype H3N2 gene NS1.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. 90

S21 **GC3 composition of the host species.** (A) Density plots of GC3 contents of avian (red), swine (green) and human (blue) hosts; vertical lines represent mean values for each density. (B) Factor effects. (C) Graphical representation of Tukey's Honestly Significant Differences (HSD). 91

S22 **GC3 composition of H1N1 viruses as a function of collection dates.** Hosts are color-coded: avian in purple, human in blue and swine in green. 92

S23 **GC3 composition of H3N2 viruses as a function of collection dates.** Hosts are color-coded: avian in purple, human in blue and swine in green. 93

S24 **Significance of the linear fit between effective number of codons and GC3 content.** H1N1 viruses are in red; H3N2 viruses are in blue. The significance level is taken at 1% (dashed horizontal line in gray). HA in H1N1 has a P-value of 0, and therefore does not appear on the graph. 94

S25 **Effective number of codons (ENC) as a function GC3 content for H1N1 viruses.** Hosts are color-coded: avian in purple, human in blue and swine in green. Gray horizontal lines represent ENC cutoffs at 35 and 50. The linear fit is represented as a black dashed line. 95

S26 **Effective number of codons (ENC) as a function GC3 content for H3N2 viruses.** Hosts are color-coded: avian in purple, human in blue and swine in green. Gray horizontal lines represent ENC cutoffs at 35 and 50. The linear fit is represented as a black dashed line. 96

S27 **Comparison of root age for H1N1 viruses with and without re-emergent strains.** The gray line represents the first bisector (line of equation $y = x$), while the red line represents the linear fit to the data. Scale on both axes is in years before 2009. Insert shows the results for all ten 'canonical' genes, with bars representing the 95% Highest Posterior Densities. 97

S28 **Cluster composition for select H1N1 and H3N2 genes.** Results of sequences clustering are shown for the fastest evolving gene, HA, for (A) H1N1 and (B) H3N2 viruses as well as for the slowest evolving gene, M1, for (C) H1N1 and (D) H3N2. Hosts are represented in red (human), green (swine) and blue (avian). Cluster IDs are arbitrary; their largest number represent the final number of sequences in the dating analyses. 98

Chapter 1

Introduction

1.1 Epidemiology of influenza

1.1.1 Pandemics and epidemics of influenza

Influenza viruses come in two epidemiologic forms of pandemic and epidemic [4]. An extensive dispersal of influenza virus on a worldwide scale is a pandemic, while an epidemic is a more sporadic and localized outbreak, as seen with seasonal influenza outbreaks. In fact, pandemic and epidemic of influenza represent a continuum of the disease.

The influenza disease cycle starts from the emergence of a novel subtype of influenza in humans (which is a pandemic caused by antigenic shift) to the circulation of closely related viruses of the same subtype in subsequent years or even sometimes several decades (which is an epidemic or a seasonal influenza caused by antigenic drift) [5]. For instance, the H1N1 viruses that circulated in humans from 1920 to 1956 were descendants of the influenza virus, which emerged in 1918 and caused the worldwide "Spanish"pandemic. However, over this period of time, the virus evolved strikingly and lost high virulence while population immunity increased against the virus. Also, epidemics that come im-

mediately after a pandemic are relatively more severe than those epidemics that occur in subsequent years or decades [5].

In addition, different factors such as epidemiological features and genetic composition of the virus can modulate the severity of a pandemic. Lack of population immunity for a new subtype of influenza virus may cause severe emergence of the virus with high attack rate and high morbidity and mortality rate [4]. Also, a mild pandemic can emerge through genetic reassortment of the virus and by acquiring internal gene segments from previously circulating influenza viruses [6].

1.1.2 Global pandemics of influenza history

During the twentieth century, three worldwide influenza pandemics have severely impacted on human population. These pandemics emerged in 1918 (Spanish pandemic, H1N1 subtype), 1957 (Asian pandemic, H2N2 subtype), and 1968 (Hong Kong pandemic, H3N2 subtype).

However, H1N1 subtype of influenza reintroduced into the human population in 1977 and led to a moderate Russian pandemic but it was much less sever than H1N1 pandemic of 1918. The reemerging H1N1 virus of 1977 did not replace the H3N2 viruses which were circulating among humans at the time and both subtypes are cocirculating among humans to this day [7].

Influenza pandemics of the twentieth century suggest that the virus emerges and spreads in different ways. The H1N1 virus of 1918 most probably originated through a cross-species transmission from an avian reservoir into an intermediate host, and further adapted virus to human host [6]. The 1957 and 1968 viruses probably originated through genetic reassortment of an avian virus to human and further adaptated virus to human host [8].

In 2009 a new H1N1 virus of swine origin triggered the first pandemic of twenty first century from the Americas. The 2009 H1N1 pandemic (swine pandemic) circulated from March 2009 according to the WHO (World Health Organisation). In fact, the 2009 H1N1 virus was generated by a triple reassortment event between swine, human, and avian influenza genes [9]. Furthermore, a direct descendant of the 1918 H1N1 virus continues circulating along with the postpandemic 1968 H3N2 reassortant virus (which had reassorted at least twice by avian influenza genes before reaching humans) and the 2009 H1N1 swine-origin pandemic influenza virus [9].

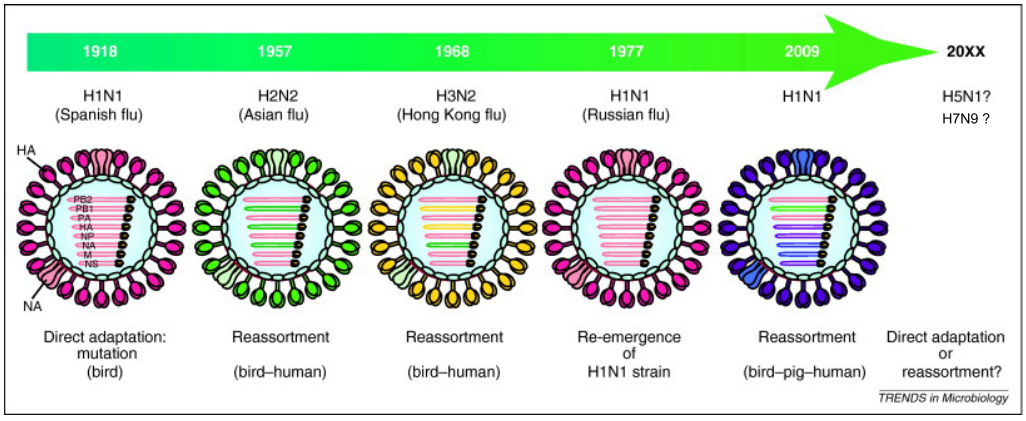

Figure 1.1: **Schematic representation of influenza A pandemics.** The past pandemics of influenza A viruses (in 1918, 1957, 1968, 1977 and 2009) are the result of cross-species transmission caused by influenza A viruses of different subtypes (H1N1, H2N2 and H3N2) through different evolutionary processes. The responsible evolutionary mechanisms for each pandemic is shown under each virus. It is likely that next pandemic would probably be a H5N1 or H7N9. Common names of the past pandemics and the virus origins are shown in parentheses at the top and bottom of each virus, respectively. [1].

1.1.3 Public health importance of influenza

Pandemic influenza is defined by the World Health Organisation (WHO) as "the most feared security threat" facing the world today. This kind of outlook puts influenza disease

away from many other diseases that may be regarded simply as medical conditions.

Also, the US pandemic plan suggests to view pandemics "as a national security issue"is critical for pandemic preparedness. While some Western governments have started to frame pandemic influenza as a threat to national security [10]. Influenza pandemic is the "highest risk"civil emergency as assessed by the "National Security Strategy of the United Kingdom". Moreover, the Australian pandemic plan led the Australian to receive one of the best possible healthcare "commensurate with the maintenance of a safe and secure society"[10].

The most devastating outbreak of influenza was the Spanish influenza in 1918 with an estimated 50 to 100 million deaths [9], while approximately 30 percent of the world's population was infected. Despite the fact that the outbreak was the era before commercial airline travels, the virus spread across the world within 2 months and killed about 5-10% of all those infected according to the National Institutes of Health (NIH). The Asian influenza pandemic in 1957 spread all over the world in just 6 months. The total death rate probably exceeded 1 million and caused roughly 70,000 deaths just in the United States according to the National Institutes of Health (NIH). The 1968 H3N2 influenza pandemic was first detected in Hong Kong, this virus caused roughly 34,000 deaths in the United States and H3N2 viruses are still around. The pandemic of the 21st century in 2009 caused approximately 274,304 hospitalization cases and 12,469 death cases only in the United States [11]. While this 2009 H1N1 virus caused 201,200 respiratory deaths along with 83,300 cardiovascular deaths globally [12].

1.1.4 Influenza transmission and required receptors

Sialic acid (Sia) serves as the receptor for primary attachment with influenza viruses. The Sia is typically combined with a galactose (Gal) in two distinct α2-3 and α2-6

configurations. These two configurations lead to a known concept of interspecies barrier (or host restriction) of influenza viruses, in which influenza viruses from different hosts preferentially bind to Sia with different Gal linkage [13]. The viral hemagglutinin (HA) binds to Sia to facilitate cellular entry of influenza virus. While the other major surface component of the influenza virus the neuraminidase(NA), cleaves Sia to release progeny viral particles from the cell [14].

The HA glycoprotein of swine influenza viruses preferentially recognizes sialic acids with both α2-3 and α2-6 -galactose linkages, while the HA glycoproteins of avian and human influenza viruses preferentially recognize sialic acids with α2-3 or α2-6 -galactose linkages, respectively [13]. The trachea of swine has sialyl receptors which attach to avian and human influenza viruses, and the trachea functions as a site for genetic reassortment of influenza viruses[15].

Unlike the 1957 or 1968 pandemic viruses, genetic analysis of the 1918 pandemic virus suggests that all of its gene segments originated from an avian reservoir without genetic reassortment [16, 17]. However, the time required for an avian-originated influenza virus to become adapted to mammals or in which mammalian reservoir, such adaptations occurred is not clear.

One common feature between the pandemic strains is acquisition of amino acid changes in the receptor binding site of the HA glycoprotein that alter the virus' receptor binding specificity from the α2-3 or α2-6 linkage between Sia and Gal [18, 19]. The switch to a predominantly α2-6 -linked sialyl receptor specificity facilitated transmission of 1918 pandemic virus and probably the 1957 and 1968 pandemic viruses [20].

The effect of the switch in receptor binding specificity on viral pathogenicity in humans is less understood. Theoretically, the changes in receptor specificity may result in change in target cells from lung epithelial cells (that exhibit α2-3 -linked receptor)

to epithelial cells lining the upper respiratory tract (that exhibit α2-6 -linked receptor), thereby reducing the occurrence of respiratory diseases such as pneumonia [6].

In addition to the surface glycoproteins, genetic analyses of influenza viruses isolated from different hosts have identified 32 residues in total from PB2, PA, NP, M1, and NS1 proteins as host-specific markers differentiating human and avian influenza viruses. Among these 32 residues, 13 residues were conserved among the 1918, 1957, and 1968 pandemic influenza viruses [21].

The clear genetic difference in these gene segments between avian and human influenza viruses, may be related to the differences in avian and human cellular machinery. It is likely that a pandemic strain should contain some of the human-specific markers that allow efficient replication and transmission [6].

Masaki Imai and colleagues (2012) showed that HA from avian H5N1 viruses can convert to an HA that supports efficient viral transmission in mammals; however, it might not be sufficient for the transmissibility of avian H5N1 virus entirely. The genetic origin of the remaining seven viral genes, which is from a 2009 pandemic H1N1 virus, might contribute to transmissibility of the virus in mammals [22].

1.2 Molecular virology of influenza

Influenza virus belongs to the Orthomyxoviridae family of viruses (from the Greek *myxa* meaning "mucus"). This virus is a medium-sized enveloped pathogen carrying a segmented negative-sense RNA as its nucleic acid.

The influenza genome has eight separate segments coding for up to fourteen different proteins. On the viral surface spike-like glycoproteins, can be identified under an electron microscope, which are hemagglutinin (HA) and neuraminidase (NA). In addition to the surface antigens, there are internal genes as bundles of segmented filaments, which are

the nucleoproteins surrounding the viral RNA.

1.2.1 Classification

Influenza viruses classify into a range of types and subtypes. There are three types of influenza viruses: A, B and C.

Influenza A viruses are capable of infecting human and a broad range of animals. High rate of evolution and adaptability to new hosts leads the influenza A viruses to be able to cause severe pandemics and epidemics in human. Type B of influenza viruses typically cause a less severe disease than type A influenza viruses. Influenza B viruses are known to infect mostly humans and seals [23]. However, influenza B viruses and influenza A viruses are morphologically and genetically very similar and both evolve through antigenic drift and reassortment [24]. Moreover, evolutionary rate of influenza B viruses is faster than C viruses and slower than A viruses relatively [25]. Type C of influenza viruses which does not cause epidemics, is more benign than viruses of type B and A and this type infects human and swine hosts[26].

Influenza A viruses are divided into several subtypes based on the properties of the surface glycoproteins HA and NA. Subtypes are characterized by eighteen different known hemagglutinins antigens (H1 to H18) and eleven different known neuraminidases antigens (N1 to N11) [27].

1.2.2 Hosts

Influenza A viruses are found in human and a range of other animal species (e.g., birds, swine, horses, minks, seals, whales, etc.) and have the capacity to cause pandemics [28]. Birds, and particularly waterfowls, are known as the natural reservoir for all influenza A subtypes [29].

The gene pool of influenza A viruses, which includes a broad range of genes from different hosts and subtypes, represents a severe contagious threat to humans. There are thus far only some varying degrees of immunity to the H1, H2 and H3 hemagglutinins and to the N1 and N2 neuraminidases [30]. In addition on rare occasions, novel viruses cross the species barriers, spilling over among hosts, and infecting people or other hosts [5].

1.2.3 Genome: structure and functions

The genome of influenza A virus consists of eight separate negative sense, single-stranded RNA segments known as viral RNA (vRNA), which can encode up to fourteen proteins. This genome has a total length of about 13,600 nucleotides.

Each segment encodes one or more viral proteins. The gene segment order of influenza A viruses is as follows: segment 1 codes for basic RNA polymerase 2 (PB2), segment 2 for basic RNA polymerase 1 (PB1) and in some strains also PB1-F2 and PB1-N40, segment 3 for acidic RNA polymerase (PA) and in some strains also PA-X [31]. Segment 4 codes for HA (a trimer of three identical subunits), segment 5 for nucleocapsid protein (NP), segment 6 for NA (a tetramer of four identical subunits), segment 7 for matrix protein1 (M1), matrix protein2 (M2 which is a tetramer of four identical subunits) and M42 [32], and segment 8 for Non-Structural protein1 (NS1) and Non-Structural protein2 (NS2) [33]. The segmented RNA genome includes the heterotrimeric ribonucleoprotein (RNP) complex (see below) coated with multiple copies of nucleocapsid protein (NP) and forms a helical hairpin structure [34]. The RNP complexes are surrounded by other genome proteins required for mRNA synthesis as well as parts of the packaging signals required during virus assembly and a layer of the matrix protein1 (M1) which is the most abundant structural protein of influenza virus [35].

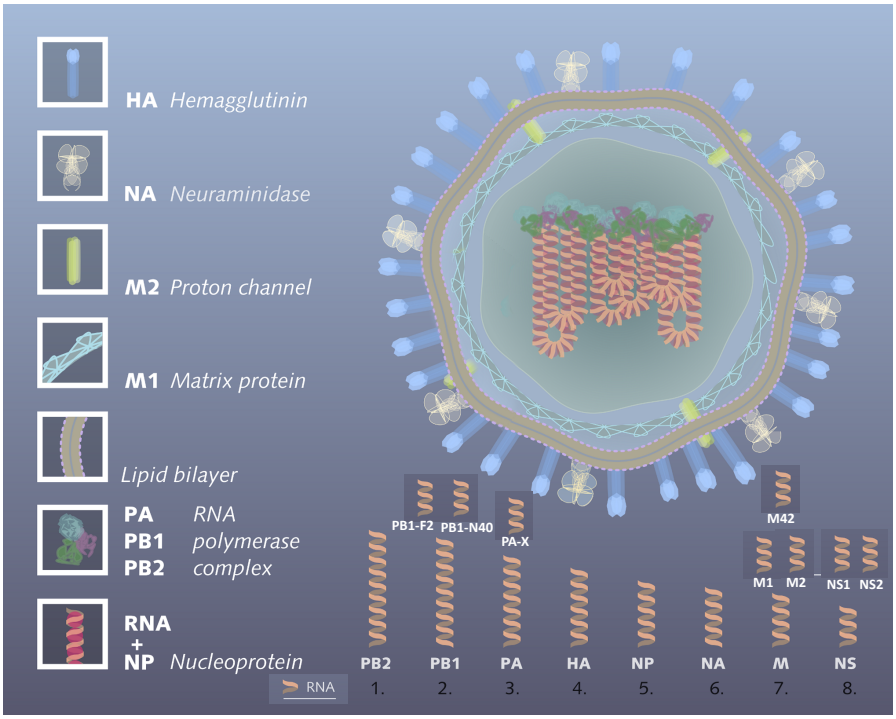

Figure 1.2: **Schematic figure of an influenza A virion.** Three proteins, HA, NA, and the M2 transmembrane proton channel, are anchored in the viral membrane, which is composed of a lipid bilayer. The large, external domains of HA and NA are the major targets for neutralizing antibodies of the host immune response. The M1 matrix protein is located below the membrane with its counterpart M42 (in some strains). Inside the virion, the eight RNA segments are packaged in a complex with nucleoprotein (NP) and the RNA polymerase complex [2]. The RNA polymerase complex also includes PB1-F2, PB1-N40 and PA-X protein segments in some strains.

(1) The RNP complex

(a) PB2, PB1, and PA The RNP complex is composed of PB2, PB1, and PA polymerases. The RNP complex which is also known as the RNA-dependent RNA polymerase complex is responsible for both transcription and replication of viral genome [33]. The complementary RNA (cRNA) is the intermediate template which is necessary for copying new viral RNA (vRNA) molecules in the nucleus. Assembly and function of the RNPs depend on complex sets of protein/protein and protein/RNA interactions [36].

The cap structures at the 5' ends of cellular mRNAs binds to PB2 and causes initiation of transcription. The capped 5' mRNA ends are then cut off by an endonuclease activity of PB1. PB1 uses this short piece of capped RNA as a primer for initializing the transcription of mRNA [37]. The helical configuration of the RNPs is determined by the structure of the NP [3]. The function of NP is to provide a structural framework for vRNA and cRNA. Also, NP might be involved in regulating viral transcription and replication and act as an elongation factor [37].

Influenza A virus segment 2 is known to encode two polypeptides in overlapping open reading frames: PB1, and PB1-F2. A third major polypeptide is synthesized from PB1 mRNA is called PB1-N40 [38].

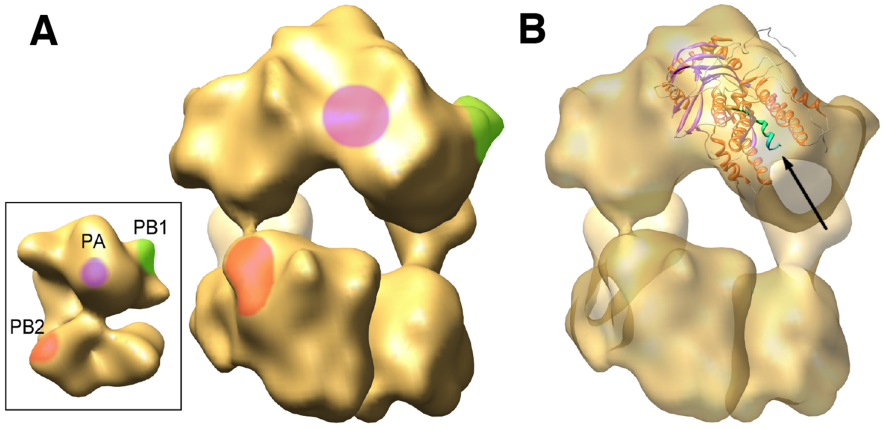

Figure 1.3: **Three-dimensional structure of the ribonucleoprotein (RNP) complex.** The schematic figure shows the three-dimensional model for: (A) the viral RNP complex including PB2, PB1 and PA. The location of specific domains in the PB1 (green), PB2 (red) and PA (violet) subunits are indicated. (B) The front-view of the polymerases is presented with the docking of the PA(C)-PB1(N) dimer and the N-terminal PB1 peptide is indicated with an arrow and highlighted in green [3].

(b) PB1-F2 The PB1 gene of influenza A viruses encodes a second small protein, PB1-F2, that is expressed from different reading frame. PB1-F2 influences the intracel-

lular localization of PB1 [38]. PB1-F2 induces apoptosis, enhances inflammation in mice, increases the frequency and severity of bacterial infections [14]. Also PB1-F2 contributes to increasing the virulence and pathogenesis of pandemic strains of influenza such as the extreme pandemic of 1918 [39].

(c) PB1-N40 The gene expression of PB1-N40, PB1-F2, and PB1 are interdependent. The PB1-N40 directs translation of an N-terminally shortened version of the polypeptide (N40) that lacks transcriptase function [38]. During the course of infection, PB1-N40 interacts with other viral proteins, therefore may have an influence on virus replication. Since PB1-N40 lacks a PA binding site, it would not be predicted to be functionally equivalent to PB1. While PB1-N40 is clearly not essential for viral growth, its absence leads to slower replication kinetics in cultured cells [38].

(d) PA-X The segment 3 of influenza genome contains two open reading frames. The first open reading frame codes for PA and the second open reading frame codes for another recently-discovered protein called PA-X [31].

The viral PA-X protein functions: (i) to inhibit cellular gene expression, (ii) to modulate viral pathogenecity in a mouse infection model, (iii) to modulate the host response to infection, or (iv) to increases in inflammatory, apoptotic and T-lymphocyte signaling pathways [31].

(2) Hemagglutinin (HA)

The major viral surface glycoprotein HA is the viral-mediated membrane fusion molecule. HA is synthesized as a precursor molecule HA0 which assembles a homotrimers. HA0 is cleaved by host enzymes into two subunits, a surface subunit (HA1) and a transmembrane subunit (HA2), activating the fusion potential of glycoproteins. The homotrimer is a

complex of 3 copies of the 2-subunit proteins (HA1 and HA2). HA1 subunit contains the receptor-binding site and is responsible for the adsorption of virus to the cell surface, while the HA2 is responsible for fusion. After binding the HA1 to its receptor, sialic acid, the influenza virus is endocytosed and fusion is initiated with the endosomal membrane at low pH [40].

(a) Receptor Binding The receptor determinant of influenza viruses is sialic acid (Sia), which binds to viral HA. Influenza virus host specificity can be explained in part by the difference in receptor-binding specificity for human, avian and swine influenza viruses [41].

The receptor-binding specificity of human, avian and swine influenza viruses suggests that avian and swine influenza viruses need to acquire the ability to recognize human-type receptors to cause a pandemic. Indeed, the earliest isolates of the 1918, 1957, and 1968 pandemics possessed avian origin HA that recognized human-type receptors. The results obtained in these studies are consistent with the theory that paucity of receptors for avian viruses in the upper respiratory tract may be one of the factors preventing efficient human-to-human transmission of avian viruses [14].

(b) Fusion Virions are internalized to host cell by receptor-mediated endocytosis, and low pH activates conformational changes in HA which is necessary for membrane fusion and initiation of infection.

The following steps can be discriminated in this process: i) The HA2 dissociates from the matrix protein M1 after proton influx into the virion via M2 ion channels. ii) The HA1 is folded back, and the HA2 is released from a cavity in the stem region behind the cleavage site and is then immersed into the endosomal target membrane. iii) The HA2 undergoes conformational changes to a hairpin structure. After formation of fusion pores

viral RNPs are delivered into the cytosol. RNPs are then transported into the nucleus, where transcription and replication occurs [14].

(3) Nucleocapsid Protein (NP)

The segment 5 of influenza viruses encodes NP, which is relatively well-conserved. NP binds to the PB1 and PB2 subunits of the RNP complexes and the matrix protein M1 directly but not with PA, both in virus-infected cells and recombinant systems [42].

The NP performs multiple essential functions throughout the virus life cycle: (i) function in encapsidation of the vRNA, which is necessary for recognition by the RNP complex, (ii) function in transporting the viral RNPs into the nucleus, in which NP interacts with the host nuclear import machinery [14], (iii) function in RNA replication (iv) vRNA synthesis, or (v) transcription [43].

(4) Neuraminidase (NA)

The segment 6 of the influenza genome encodes for the NA protein which is present in homotetrameric form in the viral envelope. The NA has several functions in the life cycle of influenza virus. The NA functions at the end of the life cycle where it enables the virus to be released from the host cell by cleaving sialic acid groups from glycoproteins.

The NA incorporates into the envelope of the budding virion and also limits influenza A virus superinfection by removing surface Sia from the virion-producing cell [44].

(5) Matrix Proteins

(a) M1 The M1 protein is the most abundant virion protein. The influenza A virus M1 protein lines the inner surface of the viral lipid bilayer [45] and interacts with the plasma membrane [46]. M1 functions in virus particle assembly, virus budding process,

interaction with cell membranes, nuclear export of RNPs later in infection [47].

(b) M2 The M2 protein of influenza A viruses is a tetrameric membrane protein which exposed at the surface of infected cells and has ion channel activity. The M2 functions in assembly of the virion, in viral replication cycle, in the release of the budding virion [48]. The M2 also maintains pH level across the viral membrane and stabilizes the virus budding site [49].

(6) Non-structural (NS) Proteins

(a) NS1 The NS1 is a small protein that regulates a number of critical events during influenza virus replication. The NS1 protein increases viral replication, enhances the rate of initiation of translation and boosts viral gene transcription by circumventing host defences [50]. The NS1 is also involved in the pathogenicity of influenza A viruses [50].

(b) NS2 In addition to the NS1, viral segment 8 encodes a polypeptide from a spliced form of the segment 8 mRNA which is called NS2 [51]. Translation of the NS2 mRNA occurs in a reading frame different from that used for NS1, thereby the NS2 gene and NS1 gene spliced out from different reading frame of the segment mRNA transcript [51]. The coding region of the NS1 mRNA overlaps the coding region of the NS2 mRNA [52] which can be additional constraints on viral codon usage bias. The NS2, also known as Nuclear Export Protein (NEP) is present in small amounts in viral particles which might interact with the viral matrix protein M1 [53]. The NS2 mediates the export of newly synthesized RNPs from the cell nucleus, thereby ensuring the viral genomic segments are available for packaging into new virions [54, 55].

1.2.4 Life cycle

Step 1: Attachment of the virus to the host cells

The influenza virus attaches to the host cell surface and initiates a long and complicated process of making more identical copies of itself. The hemagglutinin spikes on the viral surface attach to host cells at a Sia residue attached to a sugar molecule (Gal).

The precise chemical bond between the Sia and the Gal molecule determines whether influenza viruses can bind to the respiratory epithelial cells of host [5].

Step 2: Viral entry to the host cell and viral replications

Upon viral binding to the host cells, receptor-mediated endocytosis occurs and the virus enters the host cell in an endosome. The endosome has a low pH which induces the fusion of the viral and endosomal membranes and conformational changes in HA molecules [56]. Protons also enter the viral particle itself, dissociating the M1 protein and the RNP complex, such that viral genome components are free to enter the host cell cytoplasm. The viral genome components are then transported into the nucleus, where transcription and replication take place [5].

The viral genomic RNA (vRNA) functions as a template for two different RNA species: (i) complementary RNA (cRNA) and (ii) mRNA with a cap structure at the 5' end and with the 3' terminal nucleotides of cRNA being replaced by a poly-A-tail. A viral RNA dependent RNA polymerase (RdRp) initiates RNA replication of influenza viruses. The cRNA is the template for new vRNA molecules. The cap structures of the viral mRNAs are derived from cellular mRNA molecules and the viral mRNA utilizes the cellular translation machinery for the synthesis of viral proteins [14].

After that, the viral ribonucleoproteins are assembled in the nucleus and subsequently exported into the cytoplasm, while the envelope proteins are translated at the endoplas-

mic reticulum and then transported by the apparatus to the cytoplasmic membrane. From there, NA mediates virus release by removing receptors from the infected cell [56].

Step 3: Releasing new viruses which are prepared to infect new host cells

Influenza is an enveloped virus and it uses the host cellular membrane to form the viral particles that leave the cell through a budding process and go on to infect neighboring cells [56].

The final release of newly made virus is aided by the viral NA enzyme splitting off the sialic acid, which holds HA of the new virus bound to the host cell. NA here facilitates the release of new viruses. Once the newly made virus is detached from the cell, it will attempt to infect another cell [5].

1.3 Molecular evolution of influenza

1.3.1 Origin of influenza viruses

Human influenza A viruses share common ancestors with both avian and swine viruses. Although the evolutionary records are not complete, enough evidence is available to provide some insight into the origin of these viruses. Knowing the origin of the influenza viruses actually addresses two related, but distinct questions: i) in which host did the first influenza virus evolve, and ii) which host(s) are the nearest common ancestors for each of the influenza virus gene segments? [57].

Phylogenetic analyses of influenza genes, the H1 and N1 genes share common ancestors when coming from both human and classic swine virus lineages. However analysis of NP, PB2 and M genes shows that they share an avian common ancestor [57]. The common sister group relationship for all genes of human and classic swine viruses and the

closeness of the common ancestors to avian virus proteins suggest that the human and classic swine virus ancestor was not a virus affected by antigenic shift but an entirely new avian-derived virus [58]. It is important to note that the H1N1 human virus, as originally constituted prior to 1918, has since been reassorted twice: in 1957 the H1, N1, and PB1 genes were replaced with new avian genes, and in 1968 the H2 and PB1 genes were again replaced with avian genes [59].

A number of features of avian influenza viruses suggests that waterfowl might be the reservoir host of influenza viruses. The high degree of adaptation of avian viruses to their natural hosts, the considerable genetic diversity of avian virus subtypes, and the evolutionary stasis of avian virus proteins suggest that influenza viruses are a long-established pathogen in wild birds and more transient in other hosts. The evolutionary dynamics exhibited by avian viral proteins is all the more remarkable considering the geographic separation of avian virus populations and ongoing reassortment of HA, NA and other genes of avian influenza viruses [60]. During recent history there has been less isolation of human and avian virus gene pools and as a result human influenza viruses have been prevented from reaching an evolutionary equilibrium with their hosts by an irregular infusion of avian virus gene into the human virus gene pool [61].

Inferring origins and deep phylogenetic relationships of viruses is one of the most difficult issues in viral evolution to resolve. How viruses originated, and how they relate to cellular organisms, are major topics in evolutionary biology. Indeed, to some evolutionary biologists, viruses have played a central role in the establishment of complex living systems on Earth [62].

1.3.2 Genomics in studying the influenza virus evolution

High-throughput genome sequencing and antigenic typing has notably increased our understanding of influenza virus evolution. Large datasets of high-quality or complete genome sequences of viral isolates are accessible to provide insights into the pattern of world-wide spread, genetic diversity, and dynamics of subtype evolution[2].

In fact during introduction of influenza virus to humans, selection for major antigenic change of the influenza virus is the driving force in the evolutionary arms race between the virus and the immunity of the human population [63].

Major changes in viral structure and antigenicity occur when novel viruses are introduced into the human population and can be transmitted among humans [64]. Segment reassortment between different subtypes of influenza viruses of the same subtype is an important process in the evolution of human-adapted subtypes and produces extensive genome-wide diversity [65, 66].

The viral antigenic surface proteins, NA and HA, are the major targets of the host immune response [67] and following viral exposure the host immune system produces neutralizing antibodies against NA and HA.

Also antigenic comparisons of human H1, H2 and H3 influenza viruses show that each subtype has a counterpart in avian species. The H1 variants circulating in humans in the early 1930s were similar to swine viruses that were later shown to be antigenetically similar to viruses from ducks. The antigenic similarities were later confirmed by analysis of HA proteins [60].

More genomic analyses of the past pandemics show that the 1957 pandemic resulted from a reassortant H2N2 influenza virus in which both the HA and NA genes had been replaced by gene segments that are closely related to avian strains [68]. The 1968 pandemic was caused by the emergence of a H3N2 strain in which the H2 subtype of HA

gene segment was reassorted with an avian derived H3 HA gene segment, while maintaining the N2 gene segment acquired in 1957 [68, 69]. The PB1 gene segment in both the 1957 and 1968 pandemic viruses was replaced with avian counterparts [70]. The remaining five gene segments, i.e. PA, PB2, nucleoprotein, matrix and nonstructural, were all retained from the H1N1 viruses circulating before 1957 which were probably the direct descendants of the gene segments of the 1918 virus [71]. Furthermore, the genome segments of the 2009 pandemic virus derived from a complicated reassortment history with segments of: (i) avian-like Eurasian swine influenza A viruses (NA and M) which were first sampled in Eurasian swine in 1979, and (ii) A triple reassortant virus sampled from North American swine which conform from the human H3N2 (PB1), an avian influenza A virus (PA, PB2), and classical North American swine influenza A viruses (HA, NP, NS), which are descendants of the 1918 H1N1 virus [72].

1.3.3 Phylogenetic approaches to study evolution of influenza viruses

Phylogeny provides an obvious means to learn more about viral origins and evolution while phylogenetic trees represent hypotheses about evolutionary relationship among taxa. Each phylogeny of influenza gene represents a partial history of the virus, but consideration of all gene phylogenies together reveal a more complete picture of viral evolution [60].

To reconstruct phylogenies, the most common phylogenetic approaches are: (i) the neighbor-joining (NJ) algorithm and (ii) optimal tree searches that apply an optimality criterion such as parsimony or maximum likelihood (ML). The ancestor/descendant relationships between organisms or gene sequences are depicted in a graph structure that is called a phylogenetic tree [73].

The neighbor-joining reconstructs phylogenetic trees from evolutionary distant data converted into a distance matrix. This relatively very fast phylogenetic method finds pairs of operational taxonomic units (OTUs) or neighbors that minimizes the total branch length at each stage of clustering OTUs [74].

The parsimony method asserts a particular non-parametric statistical method considering all hypothetical ancestral character states with minimum evolutionary change to explain phylogenetic relationships among datasets [75].

Maximum likelihood (ML) provides estimates for a model's parameters and corrects for multiple mutational events at the same site of reconstructed tree. Reconstructing the relationships between sequences with the highest probability (the tree's likelihood) of producing the observed sequences is preferred [73].

The bayesian approach asserts a statistical model considering the posterior probability of hypotheses which is proportional to the product of the likelihood and the prior probability of the data [76]. Although Bayesian estimates are closely allied with ML but they allow complex models of sequence evolution to be implemented. Bayesian approaches have the capability of estimating divergence times, finding the residues that are important in natural selection and detecting recombination points [73].

1.3.4 The mechanisms of influenza virus evolution

Fundamental evolutionary processes of influenza viruses include mutations, natural selection, and epistasis each of which represent specific rates, determinants, and consequences carrying important role(s) in shaping evolutionary dynamics. From an evolutionary perspective, RNA viruses have two uniquely defining features: extremely high mutation rates and extremely small genomes. Furthermore, since mutations occur during nearly every round of genome replication, thus it is clear that RNA virus evolution is, to a large

extent, dominated by the process of mutation [62]. Even though there is a general consensus that replication error rates in RNA viruses are extremely high, it is important to note that the accurate estimation of these rates is challenging under any circumstance. There is also evidence that the mechanism of viral replication affects mutation rate. Moreover, in fact precisely describing the mutation distribution is critical to understanding adaptation. For example, the occurrence of multiple advantageous mutations in a single replication cycle may be critical for successful cross-species virus transmission [62].

1.3.5 Major evolutionary changes of influenza viruses

Influenza A viruses undergo two types of change affecting their major surface glycoproteins called antigenic shift and antigenic drift. Since the start of modern virology there have been antigenic shifts that occurred in 1918, 1957, and 1968. A shift that results in a pandemic (a worldwide epidemic), is always preceded by an abrupt change in hemagglutinin subtype and, apart from the 1968 virus, also by a change in neuraminidase subtype. Drift results in epidemics, and is caused by gradual evolution under selective pressure of neutralizing antibodies. A new shift virus immediately starts to undergo continuous antigenic drift [77]. The complex evolutionary processes that influenza viruses undergo can be better understood against a background of their biology. These viruses are highly infectious and cause an acutely cytopathogenic infection. Thus, they are a victim of their own success which results in a near universal immunity among hosts who, given adequate conditions, can live for a long time. The phenomena of antigenic shift and drift in influenza have significant implications for human infections and they are monitored by a worldwide network of laboratories, coordinated by the WHO, which isolate and classify currently circulating influenza viruses. In this way new strains can be quickly spotted and an appropriate vaccine can be developed in a timely manner [77].

Mechanism of antigenic shift Antigenic shift has occurred sporadically from the first recorded shift of the twentieth century in 1918. Antigenic shift is the introduction of a virus into a population that has no pre-existing immunity, so it is mostly associated with a pandemic with high morbidity and mortality, although these vary between different shifts. Since 2001, a new H1N2 reassortant virus has been isolated in many countries in mainly young people. However, this H1N2 has not spread widely, presumably because it is sensitive to existing immunity to its parent human H1N1 and human H3N2 viruses [77]. Until 1977, only one virus subtype was in circulation at any one time, and this virus was replaced completely when a new subtype was introduced. What causes the original subtype to disappear is not known. In 1977, during the "reign"of the H3N2 subtype, an H1N1 virus appeared that was identical to the H1N1 virus that had been in circulation from 1918. Thus this was not strictly speaking a shift, but a reintroduction or a lab escape of H1N1. The 1977 virus had not been infecting the human population during its "absence"as it would have undergone 27 years of antigenic drift. The drifted descendants of the 1968 H3N2 and the 1977 H1N1 cocirculate and continue to undergo antigenic drift. However a new shift could occur at any time [77]. As indicated, the natural reservoir of influenza A viruses comprises wild aquatic birds, such as ducks, terns, and shore birds. However, like most viruses, avian influenza virus is species-specific and does not readily infect other bird species let alone mammals, which are members of another major taxonomic group. Thus, the virus has to adapt by progressive mutations before the virus can establish itself into human hosts [77].

One of the major gaps in modern virology is the understanding of animal virus transmission in general, and what gene or genes are responsible for its control. Another variation in the evolution of virus from wild birds to humans may involve domestic pigs. In rural areas poultry and pigs are often kept together, giving ample opportunity for the

crucial bird-to-mammal adaptation. At this point the virus has two evolutionary options. Virus can continue to accumulate mutations and become better adapted to humans, or it can reassort with a human influenza virus strain and so acquire genes that are already fully adapted [77].

Mechanism of antigenic drift Influenza A viruses have been isolated every year from humans around the world since their discovery in 1933. Each new isolate is tested serologically with antibody to all other influenza strains. It soon became apparent that antigenic drift is explained by influenza virus mutants carrying modified antigenic determinants have an evolutionary advantage over the parent virus in the face of the existing immune response. Thus, year by year, new amino acid substitutions and new isolates appear which are antigenically slightly different from those of the previous year. Moreover, antigenic drift is as important in causing human influenza as antigenic shift. Thus, in the twentieth century, drift viruses were responsible for approximately 40 to 140 million deaths worldwide [77]. Antigenic drift begins as soon as a new shift virus appears and infects people. Drift of influenza A viruses is linear due to the dominating effects of favorable mutations. Drift happens on a global scale. It can only be theorized how drift takes place as it is an assumption that drift variants arise from virus circulating in the previous year. It is generally believed that variants are selected by neutralizing antibody [77].

1.3.6 Genetic changes

Molecular genetic evidence

Codon usage bias is a distinctive characteristic of many organisms and has been noted in viruses such as influenza. As the influenza virus relies on the host cell machinery

for its replication, codon usage of its viral genes might be subject to host selective pressures and adaptation especially after interspecies transmission [78]. Viral codon usage is shaped by the conflicting forces of mutational pressure and selection to match host patterns for optimal expression [79]. A better understanding of viral evolution and host adaptive responses might help control this disease. Codon usage patterns allow identification of host origin and evolutionary trends in influenza viruses, providing an alternative method and a tool to understand the evolution of influenza viruses. Human influenza viruses are subject to selection pressure on codon usage, which might assist in understanding the characteristics of newly emerging viruses [78]. Codon usage bias, which is largely determined by the nucleotide composition in the third codon position, allows a different perspective on viral evolution to be examined. Studies on codon usage bias retain some of the underlying structure of the coding sequences and may give another perspective on evolutionary changes [78]. However, little has been done to investigate the effect of selection pressure imposed by the human host on the codon usage of human influenza viruses and trends in viral codon usage over time. Fortunately, many different methods of codon usage bias detection have been developed [80]. Codon usage bias indexes include GC content, effective number of codons (ENC), Synonymous Codon Usage Orders (SCUO), Codon Volatility, Relative Synonymous Codon Usage (RSCU), and Odds Ratio [81]. Fancher and colleagues [81] have revealed that for influenza A viruses, a RSCU value is positively correlated to the Odds Ratio within that codon. The GC content, ENC, SCUO, and Codon Volatility show same pattern for the avian, human, and swine hosts; however, the RSCU and Odds Ratio of the hosts are not similar [81]. Here, we focus on two frequently used bias indexes: GC content and ENC which they have been widely used in previous works [78, 82, 83]. The frequency of which a Guanine (G) or a Cytosine (C) nucleotide appears at the third position of the codons in

a gene is the GC3 content (GC content of the third codon position). This measurement has been shown by previous studies to correlate very strongly with the codon usage bias of a gene [81]. The GC content provides a simple proxy about other codon usage bias indices because of its strong correlation with the usage bias on the entire gene [83, 81]. The ENC is another simple metric used to quantify the synonymous codon usage bias of a gene. ENC estimates the absolute synonymous codon usage bias, which will range from 20, when only one codon is used per amino acid, to 61, when all synonymous codons are used with equal frequency. The ENC analysis yields an easy-to-understand representational value for the synonymous codon dispersion within a gene. However, ENC method is still quite limited in that it does not provide specific details on codon usage frequency [81, 84].

1.4 Objectives of the thesis

Remarkable developments have emerged within recent years in the dynamic of evolution and phylogenetic studies of influenza viruses. However, the knowledge in this field is still limited, which justifies the need for more systematic efforts. The main objective of this thesis is to study the adjustment dynamics of influenza A viruses to a new host. In particular, to unravel (i) if some of the genes making up the influenza genome are more critical than others when adjusting to a new host, (ii) if this process depends on the direction of host change, or (iii) if there is some variation among influenza A subtypes. To address these questions, we developed a new procedure to estimate the dynamics of the emergence of stable influenza A lineages following a cross-species transmission. This thesis (except for Chapter 1 and chapter 4) is a reproduction of our PLoS One 2013 publication [85]. Based on a phylogenetic approach, we reconstructed the history of both host and GC3 changes in the two most human-prevalent influenza A subtypes, H1N1

and H3N2, focusing on three hosts in which both of these subtypes have established themselves: human, avian and swine. With the analysis of almost 100 years of complete genomes collected in North America, we show that two genes, NA and NS2, adjust to a new host relatively quickly. We also show that the adjustment process is asymmetric among hosts, with viruses of avian origin adjusting the fastest. Finally, while the ranking of fast-adjusting genes is the same for both H1N1 and H3N2 subtypes, selective constraints of H3N2 are relaxed faster than for H1N1 viruses.

Chapter 2

Methods

2.1 Data collection and alignment

Whole genome sequences of H1N1 and H3N2 subtypes of *all* influenza A viruses collected between 1900 and 2009 (as of January 2010) in North America (Mexico, the USA and Canada) in avian, human and swine hosts were retrieved from the Influenza Virus Resource [86]. Only one pandemic H1N1/2009 genome was included in this study, A/Canada-AB/RV1531/2009(H1N1) ; A/Saskatchewan/5131/2009(H1N1) is a seasonal (pre-pandemic) H1N1 virus [87].

The complete influenza genome includes the ten 'canonical' protein-coding genes [88, 89], consisting of the three polymerase subunits PB2, PB1 and PA, the hemagglutinin (HA) and neuraminidase (NA) antigens, the nucleoprotein (NP), ribonucleoprotein exporter (NS2, also called NEP), interferon antagonist (NS1), ion channel protein (M2) and the matrix protein (M1). Each gene was aligned at the protein level with `Muscle` [90] and back-translated to nucleotide alignments with `Pal2Nal` [91]. At this stage, manual adjustments were performed, in particular for the M2, M1, NS2 and NS1 genes. Improperly annotated or misaligned sequences were discarded. In total, our initial alignments

Table 2.1: **Genomic data used in this study.** *sl*: sequence length (nucleotides); ns_{total}: number of sequences before clustering; ns_{clust}: number of sequences after clustering with `DOTUR`.

Subtype	Gene	*sl*	ns_{total}	ns_{clust}
H1N1	PB2	2277	1900	81
	PB1	2271	1932	82
	PA	2148	1931	75
	HA	1698	1921	87
	NP	1494	1923	77
	NA	1407	1910	79
	M2	291	1906	59
	M1	756	1906	66
	NS2	363	1915	66
	NS1	657	1915	80
H3N2	PB2	2277	1107	45
	PB1	2271	1110	45
	PA	2148	1088	36
	HA	1698	1106	59
	NP	1494	1064	39
	NA	1407	1015	55
	M2	291	907	32
	M1	756	907	29
	NS2	363	1104	48
	NS1	657	1090	38

contained 19,159 H1N1 and 10,498 H3N2 genes (Table 2.1).

2.2 Sequence clustering

In order to decrease sample size to make alignments amenable to phylogenetic analysis without compromising data quality, sequences similar at the 99% threshold were removed from the alignment as done in a previous study [89]. Briefly, pairwise genetic distances were computed with `PAUP*` [92] under the GTR + Γ + I model of evolution. Sequences were then clustered with `DOTUR` [93] at the 99% similarity level using the nearest neighbor

algorithm. We checked that each cluster thus identified contained sequences coming from only one single host (Fig. S28); when this was not the case, a sequence from the most common host was selected at random; we then tested that such cases correspond to unsustained cross-species transmission events (Fig. S1-S20), so that these cases are not included in our dating analyses. Note that the H1N1 1918 human virus [94] was not included in the final data. Accession numbers of the genes retained are shown in Fig. S1-S20.

2.3 Phylogenetic analyses

The most appropriate model of evolution for each of the ten 'canonical' gene of each subtype was chosen according to the Akaike Information Criterion in `jModelTest` [95] (Table 2.2).

The strict molecular clock was tested with `PAML` ver. 4.4b [96] under the TipDate model [97] using the trees estimated under a relaxed molecular clock implemented in `BEAST` ver. 1.6.1 [98].

Divergence times were estimated by assuming an uncorrelated lognormal prior distribution to describe the evolution of the rates of evolution [99]. A Bayesian coalescent skyline prior with ten breakpoints and stepwise splines [100] was placed on times. Markov chain Monte Carlo samplers were run for 1 billion steps with a thinning of 5000 steps for each gene, and in duplicate to check for convergence. `Tracer` (`tree.bio.ed.ac.uk/software`) was used to monitor the runs and to determine the burn-in periods. An in-house Perl script was then used to remove the burn-in period of each pair of runs, concatenate the log files and run `TreeAnnotator` [98]. The relaxed-clock trees are, by construction, rooted (*e.g.*, [101, 89]).

Table 2.2: **Best-fit substitution models of evolution.** Models were selected based on the Akaike Information Criterion (AIC).

Subtype	Gene	AIC-selected model	Closest model in `BEAST`
H1N1	PB2	GTR + Γ_4 + I	GTR + Γ_4 + I
	PB1	GTR+ Γ_4	GTR + Γ_4
	PA	GTR + Γ_4 + I	GTR + Γ_4 + I
	HA	GTR + Γ_4 + I	GTR + Γ_4 + I
	NP	GTR + Γ_4 + I	GTR + Γ_4 + I
	NA	GTR + Γ_4	GTR + Γ_4
	M2	TPM2uf + Γ_4	HKY+ Γ_4
	M1	TVM + Γ_4 + I	GTR + Γ_4 + I
	NS2	TPM1uf + Γ_4	HKY+ Γ_4
	NS1	TVM + Γ_4 + I	GTR + Γ_4 + I
H3N2	PB2	GTR + Γ_4 + I	GTR + Γ_4 + I
	PB1	GTR+ Γ_4	GTR + Γ_4
	PA	TVM + Γ_4	GTR + Γ_4
	HA	GTR + I	GTR + I
	NP	TVM + Γ_4	GTR + Γ_4
	NA	GTR + I	GTR + I
	M2	HKY + I	HKY + I
	M1	GTR + I	GTR + I
	NS2	TVM + I	HKY + I
	NS1	TVM + I	GTR + Γ_4 + I

2.4 Timing GC3 adjustment after a host change

Host changes were determined by mapping ancestral hosts on the phylogeny of each gene under a simple maximum likelihood approach [102, 103] assuming that all three hosts had the same rate of change (more sophisticated models where all rates were different tended to exhibit convergence issues on our data). The `APE` library [104] in `R` [105] was used for this purpose. Placement of host-switch events was determined manually according to reconstructed ancestral mapping.

GC3 content and effective number of codons (ENC) were calculated for each gene with `GCUA` [106]. Gene-specific GC3 distributions were discretized by Partition Around

Medoids clustering, where the optimal number of clusters was determined by Median Split Silhouettes (for details, see [107]). Ancestral GC3 cluster assignments were reconstructed with a maximum likelihood model as above [102, 103]. Stabilization of GC3 content was inferred when (i) a host change occurred along a lineage and (ii) a subsequent change of GC3 cluster occurred. Because of the uncertain ancestral reconstructions for the two branches emanating from the root, these two branches were left out of the computations. Adjustment times were inferred as depicted in Fig. 3.2.

Table 2.3: **Significance of the trend observed in GC3 clusters.** n_{MSS} is the number of clusters estimated by median split silhouettes. P-values are derived from robust linear regressions. Significant results ($\alpha < 0.01$) are in bold. Cluster identifiers are arbitrary.

Subtype	Gene	n_{MSS}	GC3 cluster 1	GC3 cluster 2
H1N1	PB2	2	**0.0002**	**9.508**$\times 10^{-8}$
	PB1	2	0.7029	0.3913
	PA	2	0.1839	0.5605
	HA	3	0.0115	0.6235
	NP	2	0.4107	**0.0001**
	NA	2	0.8631	**0.0079**
	M2	9	0.1014	0.2161
	M1	9	0.9716	0.8673
	NS2	2	0.7784	0.4993
	NS1	2	0.0666	0.9948
H3N2	PB2	2	**0.0002**	**0.0056**
	PB1	2	0.3329	0.9434
	PA	2	0.2945	0.0762
	HA	2	0.4907	**4.828**$\times 10^{-5}$
	NP	2	0.8473	0.1514
	NA	2	0.2762	0.4825
	M2	9	0.0238	0.1936
	M1	2	0.6171	0.8301
	NS2	2	0.8538	0.0274
	NS1	7	0.5675	0.8012

We also downloaded from *ensembl* release 62 [108], available at `ensembl.org/info/data/ftp`, the complete transcriptomes of the hosts: chicken (*Gallus gallus* – chosen

arbitrarily out of the three completed bird genomes with turkey and zebra finch, as of October 2011), human (*Homo sapiens*) and pig (*Sus scrofa*). The transcriptomes were analyzed with `GCUA` and tested for transcriptomes-wide differences in their GC3 composition. Genes with no termination signal as *per* `GCUA` or with $> 20{,}000$ bases were discarded, leaving 17,087 avian genes, 46,040 human genes and 14,056 swine genes.

2.5 Detection of selection

In order to test for positive selection at some sites along the branches between a host change and a change of GC3 cluster, we ran branch-site codon models [109] as implemented in `codeml` ver. 4.4d [96]. Nonsynonymous to synonymous rate ratios (ω) are used to measure selection in protein-coding genes, with $\omega < 1$ indicating negative selection, $\omega = 1$ neutral evolution and $\omega > 1$ positive selection. Branch-site codon models allow ω to vary both along the sequence and along some pre-specified branches, called the foreground branches, while the ratio in the other branches, or background branches, is kept constant and < 1. A likelihood ratio test (LRT) was used to test the null hypothesis H_0 that there is no positive selection at any site along the foreground branches. The alternative H_1 is that there is evidence for positive selection at some sites in the foreground branches. The LRT test statistic was conservatively assumed to follow a χ^2 distribution with one degree of freedom rather than the appropriate mixture distribution [109].

Sites potentially evolving adaptively were inferred with a Bayes empirical Bayes method [110] at the 95% posterior probability cutoff. All regressions performed in this study were based on robust linear models [111].

Table 2.4: **Branch-site test of positive selection during adjustment to a new host**. Hyp: hypothesis; ℓ: log-likelihood; np: number of parameters; P: P-value; $\hat{\omega}$: estimated ω value; $\hat{p}_\omega$: proportion of sites (with $\hat{\omega} < 1$ under H_0, or $\hat{\omega} > 1$ under H_1); na: not applicable; long dash (—): data not available.

Subtype	Gene	Hyp	ℓ	np	P	$\hat{\omega}$	$\hat{p}_\omega$	Sites
H1N1	PB2	H_0	-20430.45	164	na	0.03	0.97	na
		H_1	-20430.45	165	1.0000	1.00	0.00	none
	PB1	H_0	-20277.40	166	na	0.03	0.96	na
		H_1	-20277.40	167	0.9980	1.00	0.03	75 E, 327 R, 741 A
	PA	H_0	-19132.36	152	na	0.03	0.96	na
		H_1	-19132.36	153	0.9917	1.00	0.00	none
	HA	H_0	-17467.62	176	na	0.06	0.85	na
		H_1	-17467.32	177	0.4409	2.08	0.00	277 G
	NP	H_0	-12072.00	156	na	0.03	0.93	na
		H_1	-12072.00	157	1.0000	1.00	0.03	none
	NA	H_0	-12965.21	160	na	0.06	0.82	na
		H_1	-12965.21	161	0.9952	1.00	0.00	none
	M2	H_0	-1733.00	120	na	0.07	0.53	na
		H_1	-1733.00	121	1.0000	1.00	0.00	none
	M1	H_0	-4747.81	134	na	0.02	0.97	na
		H_1	-4747.81	135	1.0000	1.00	0.01	none
	NS2	H_0	-2538.76	134	na	0.04	0.80	na
		H_1	-2538.61	135	0.5795	1.56	0.02	none
	NS1	H_0	-6083.67	162	na	0.12	0.83	na
		H_1	-6083.67	163	1.0000	1.00	0.00	none
H3N2	PB2	H_0	-12423.36	92	na	0.03	0.95	na
		H_1	-12423.21	93	0.5828	1.32	0.02	65 E
	PB1	H_0	—	—	na	—	—	na
		H_1	—	—	—	—	—	na
	PA	H_0	-9760.82	74	na	0.03	0.96	na
		H_1	-9760.82	75	1.0000	1.00	0.02	388 S
	HA	H_0	—	—	na	—	—	na
		H_1	—	—	—	—	—	na
	NP	H_0	—	—	na	—	—	na
		H_1	—	—	—	—	—	na
	NA	H_0	-8058.08	112	na	0.05	0.66	na
		H_1	-8058.06	113	0.8468	1.63	0.09	none
	M2	H_0	—	—	na	—	—	na
		H_1	—	—	—	—	—	na
	M1	H_0	—	—	na	—	—	na
		H_1	—	—	—	—	—	na
	NS2	H_0	-1936.06	98	na	0.05	0.81	na
		H_1	-1935.31	99	0.2203	2.22	0.04	none
	NS1	H_0	-3412.36	76	na	0.09	0.72	na
		H_1	-3412.36	77	1.0000	1.00	0.00	none

Chapter 3

Results

3.1 Sequence clustering

In order to estimate viral adjustment times in influenza A viruses after a host change, we retrieved the sequences of *complete* genomes of H1N1 and H3N2 subtypes from the Influenza Virus Resource [86]. We specifically downloaded *all* the genomes collected in North America (Mexico, the USA and Canada) between 1900 and 2009. Only one pandemic H1N1/2009 genome was included in this study [87]. This lead to an average of 1916 H1N1 and 1050 H3N2 sequences per gene.

After alignment, the size of the data sets was reduced to make them amendable to the Bayesian relaxed molecular clock analyses. Pairwise genetic distances were computed and clustered with the nearest neighbor algorithm; clusters of sequences similar at the 99% level were formed and a sequence representative of each cluster was drawn (see Methods for details and constraints). This clustering reduced the size of the data to more manageable numbers with an average of 75 H1N1 and 43 H3N2 sequences (Table 2.1). These data sets therefore stand as representative samples of the exhaustive whole-genome diversity deposited in GenBank (as of January 2010). This reduction step

affects the hypothesis underlying the coalescent process used as a prior distribution in the estimation of divergence times used below. However, since we (i) did not attempt to reconstruct ancestral demographics (viral incidence), (ii) used the same process to analyze both subtypes and (iii) expected that most adjustment periods did not occur following recent host changes, this reduction step is unlikely to bias the comparison of adjustment dynamics of H1N1 and H3N2 viruses.

3.2 H1N1 and H3N2 subtypes evolve with extensive reassortment

Under this general framework, we reconstructed dated phylogenetic trees for all ten 'canonical' protein-coding genes of influenza viruses [88, 89] of the H1N1 (Fig. S1-S10) and H3N2 subtypes (Fig. S11-S20) under a relaxed molecular clock [99]. Note that we assumed a single (time-homogeneous) model of evolution instead of using nonhomogeneous models [112]; this choice could potentially impact the estimated trees, but a number of empirical studies have now shown that this concern may not be warranted (*e.g.*, [113, 114]). Because the natural host of influenza viruses is considered to be avian [115], we expected that bird viruses would diverge first in all estimated trees. We also expected to find similar phylogenies for all ten genes within a given subtype, as the data come from the same individual viruses. However, the trees estimated here show a variety of scenarios, all with a posterior probability of 1 at the root node. Only PB2 and PA consistently show an avian-first split across the two subtypes, along with NP in H1N1 and NA and NS2 in H3N2 subtypes (Fig. S1-S20). Known reassortment events are also recovered here, as in the case of A/Saskatchewan/5131/2009(H1N1), one of the two "H1N1_Canada_Human_2009" genomes in Fig. S1-S10, which is a reassortant virus for

which: (i) HA and NA are derived from the non-pandemic A/Brisbane/59/2007 human virus, as seen in Fig. S4 and S6, (ii) PB2, PB1, PA, NP, M and NS are of swine origin (Fig. S1-S3, S5 and S7-S10) and (iii) that this virus emerged during the late 1990's; all these results are consistent with the original study [87], which therefore suggests that our results are not data-dependent. These results nonetheless highlight that extensive amounts of reassortment (exchange of RNA segments between viruses) exist, at least within each subtype.

3.3 Some genes evolve faster in H3N2 than in H1N1

A by-product of the relaxed molecular clock models used here is the estimation of gene-specific absolute rates of evolution. Figure 3.1 shows that these rates are systematically larger for H3N2 than for H1N1 viruses, with a genome-wide average of 2.38×10^{-3} (SEM $= 1.48 \times 10^{-4}$) and 2.01×10^{-3} (SEM $= 2.33 \times 10^{-4}$) substitutions/site/year, respectively, but not significantly so (test on the intercept: $t_8 = 0.67$, $P = 0.5245$).

These estimates are very close to those previously reported [116] or with earlier knowledge of relative rates of evolution of H3N2 and H1N1 viruses [117]. The rate difference between the two subtypes appears to be significant (at the 5% level) only for three genes (HA, NA and NS2; Fig. 3.1). Because H3N2 has been the dominant subtype in human populations for the 40 years preceding 2009, it can be posited that these genes are under stronger selective pressure than in H1N1 subtypes.

The most salient feature of Fig. 3.1 is the linear relationship, on a $\log_e$-$\log_e$ scale, between the gene-specific rates of evolution of H3N2 and H1N1 subtypes ($P = 6.8 \times 10^{-5}$; $R^2 = 0.66$) which indicates that the fast-evolving genes are the same in both subtypes. The simplest explanation, mechanistic in nature, would be that each gene accumulates substitutions at a gene-within-subtype specific rate, that is, follows a strict molecular

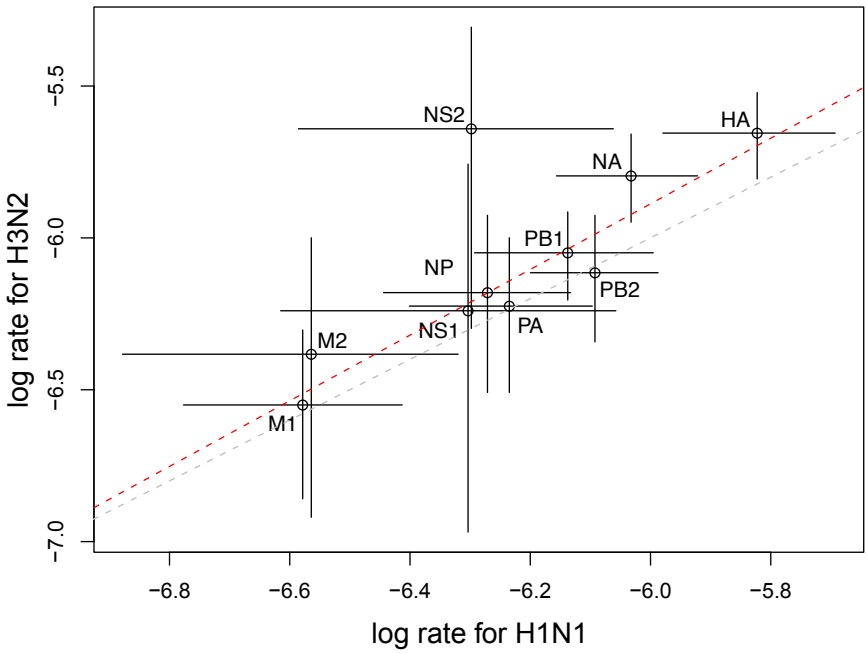

Figure 3.1: **Posterior mean rates of evolution of H3N2** *vs.* **H1N1 viruses.** Results are shown on a $\log_e$-$\log_e$ scale (in substitutions/site/year). The gray line represents the first bisector (line of equation $y = x$), while the red line represents the linear fit to the data. Bars: limits of the 95% Highest Posterior Densities.

clock [118]. However, this hypothesis is strongly rejected (Table 3.1).

An alternative explanation is that the fast-evolving genes (HA and NA) are expressed at the surface of the viral particle and are directly involved in the immune escape of the virus, while the slow-evolving genes all have internal functions [116]. NS2, which is also a fast-evolving protein, interacts directly with a host protein [119] and might therefore be involved in an 'arms race' with the host, leading up to high rates of evolution.

3.4 Estimation of GC3 adjustment times

All the results above are consistent with previous reports, but they do not inform us on the time it takes for a virus to adjust to a new host. We define this duration by the

Table 3.1: **Tests of the molecular clock assumption**. ℓ_{noclock}: log-likelihood without the clock assumption; ℓ_{clock}: log-likelihood under the clock assumption; ns: number of sequences; X^2: test statistic (twice the log-likelihood difference); df: degree of freedom; P: P-value.

Subtype	Gene	ℓ_{noclock}	ℓ_{clock}	ns	X^2	df	P
H1N1	PB2	-21268.37	-21497.97	81	459.21	79	5.604×10^{-55}
	PB1	-21193.70	-21385.71	82	384.03	80	2.801×10^{-41}
	PA	-19708.20	-19957.45	75	498.50	73	1.282×10^{-64}
	HA	-17590.60	-17828.70	87	476.21	85	9.481×10^{-56}
	NP	-12609.96	-12761.42	77	302.91	75	3.708×10^{-29}
	NA	-12671.49	-12814.82	79	286.65	77	6.612×10^{-26}
	M2	-1681.90	-1727.97	59	92.14	57	0.0022
	M1	-4855.03	-4909.53	66	109.01	64	0.0004
	NS2	-2585.38	-2721.16	66	271.56	64	2.221×10^{-27}
	NS1	-6115.67	-6217.65	80	203.95	78	3.277×10^{-13}
H3N2	PB2	-13255.71	-13331.62	45	151.81	43	4.673×10^{-14}
	PB1	-11495.45	-11558.25	45	125.59	43	5.107×10^{-10}
	PA	-10460.50	-10518.00	36	116.99	34	4.844×10^{-11}
	HA	-10919.79	-11000.53	59	161.46	57	6.317×10^{-12}
	NP	-7729.46	-7789.52	39	120.11	37	1.032×10^{-10}
	NA	-8360.71	-8429.88	55	138.34	53	1.501×10^{-09}
	M2	-1254.13	-1290.31	32	72.35	30	2.339×10^{-05}
	M1	-2805.23	-2826.02	29	41.58	27	0.0362
	NS2	-2006.66	-2155.28	48	297.24	46	1.816×10^{-38}
	NS1	-3478.96	-3519.68	37	81.44	35	1.443×10^{-05}

period delimited by two events: a host change, followed by a change of viral GC3 content in the new host. Host changes were mapped using a simple maximum likelihood model [102, 103] on the phylogenetic trees estimated above. To ease computations, observed GC3 compositions were discretized (clustered) and, just like host changes, mapped on the estimated phylogenetic trees. This process was repeated for each gene of the influenza A genome, in each subtype.

Four remarks are necessary at this point. First, GC3 content is often used to monitor viral codon optimization after a host change, as in HIV-1 [120] and bacteriophages [121].

Furthermore, codon usage has been shown to be host-specific in the case of influenza viruses [78]. Here for instance, three human and swine data points in PB2 of H1N1 are in the avian GC3 cluster (Fig. S22), and the phylogenetic analysis clearly demonstrates their recent avian origin (Fig. S1). Yet, a change in GC3 composition does not necessarily reflect an adaptive process (see below). A critical asset of our computational approach is that we do not assume any adaptive process. Second, this process of GC3 change following cross-species transmission is obviously gradual. Similar processes have been documented both experimentally in HIV-1 [120] and computationally in bacteriophages [121], and no evidence ever suggested any form of stepwise (instantaneous) adjustment. Our discretization of the process can therefore be seen as a heuristic, but one that makes the computation more straightforward than fitting a diffusion process and determining the point at which *e.g.* 95% of the GC3 content has reached a new stationary phase. Third, an alternative to reconstructing changes of GC3 clusters would have been to reconstruct the sequences of ancestral genomes in order to compute GC3 contents on these ancestral genomes. However, while accuracy of ancestral sequence reconstruction can be high ($> 90\%$) with four amino acid sequences [122], the actual performance of these methods with dozens of DNA sequences is unknown. Although ancestral state reconstruction might be more powerful, we opted here to reconstruct changes of GC3 contents directly. Fourth, phylogenetic uncertainty could be taken into account in our reconstructions of both host and GC3 changes, for instance by running the algorithm on all the trees sampled from the posterior distribution. We did not attempt to perform this computationally demanding analysis, as the objective here essentially aims at demonstrating the feasibility of the approach.

While we can estimate the dates beginning and terminating a branch on which each event (host-switch, GC3 cluster change) occurred, we do not know the exact time when

each event took place. Nonetheless, we can define two durations, a maximum and a minimum duration indicated as $\max_t$ and $\min_t$, respectively, as in Fig. 3.2. The estimated adjustment periods used henceforth are the arithmetic averages of $\max_t$ and $\min_t$.

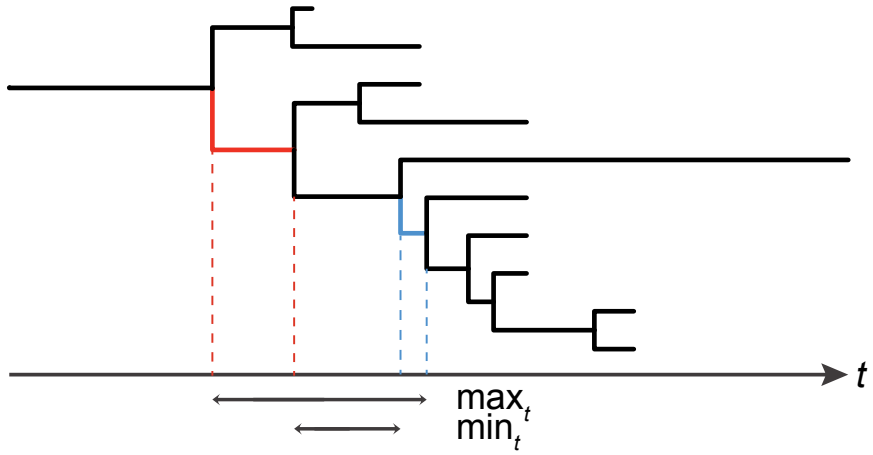

Figure 3.2: **Estimation of adjustment times.** Schematic representation of the method developed to estimate adjustment times. A host-switch event occurred along the red branch, and a GC3 cluster change occurred along the blue branch. Time t flows from the past to the present (bottom axis), and divergence times are estimated for nodes (see vertical broken lines). The two durations of interest are $\max_t$ and $\min_t$. See text for details.

3.5 GC3 adjustment is faster in H3N2 than in H1N1

The GC3 adjustment process following a host change implicitly assumes that all three hosts have different GC3 compositions, and that the GC3 content of viruses tends to reflect that of their host. We detected a significant difference in the GC3 compositions of the transcriptome of all three hosts ($F_{2,77179} = 862.81$, $P < 2.2 \times 10^{-16}$), with birds having the largest GC3 content, followed by swine (Fig. S21). Notably, GC3 contents of influenza viruses coming from specific hosts are ranked in the same order (Fig. S22-S23), but tend to be twice as high as those of their host. This might explain why decreasing

GC3 trends have been observed within host-specific influenza viruses (*e.g.*, [78]). We observed such trends here, but most of them were not significant, even using robust regressions (Table 2.3). This approximate stationarity of within-host GC3 contents gives further ground to our discretizing them. Indeed, our assumption of the existence of host-specific viral GC3 content demands that GC3 content be approximately constant in time. If this were not the case, we would not be able to draw horizontal lines in Fig. S22 and S23 to represent boundaries between these host-specific GC3 contents.

GC3 compositions of each gene of both H1N1 and H3N2 clustered into two groups (as determined by median split silhouettes; Fig. S22-S23, Table 2.3) for most genes, typically clustering human and swine hosts together. In order to simplify the algorithm, we forced clustering to have two groups for each gene. As above for rates, we found a ($\log_e$-$\log_e$) linear relationship in terms of GC3 adjustment durations between the two subtypes (Fig. 3.3). In particular, (i) GC3 content of H3N2 viruses adjusts faster than in H1N1 viruses ($F_{4,5} = 67.65$, $P = 0.0012$) and (ii) the same ordering of genes exists for both subtypes ($P = 0.0006$; $R^2 = 0.96$). It is interesting to note that genes that are fast adjusting are also involved in the final stages of the viral cycle, NS2 mediating the export of newly synthesized RNPs from the nucleus [55] and NA mediating virus release from the infected cell [56]. Note that some genes are missing from Fig. 3.3 because they did not show any evidence for a combined host/GC3 change in our genome catchment. These are H3N2 genes PB1, NP, M2 and M1. Ordering for these genes in both subtypes was achieved by fitting a linear model (ANOVA) that describes mean GC3 change times as a function of two factors: gene segment and direction of host change. Results show that these two factors have a very significant effect ($P = 1.25 \times 10^{-6}$ and 5.74×10^{-10}, respectively; Fig. 3.4), so that three points can be made.

First, the rank ordering of genes by their adjustment time differs from their ranking

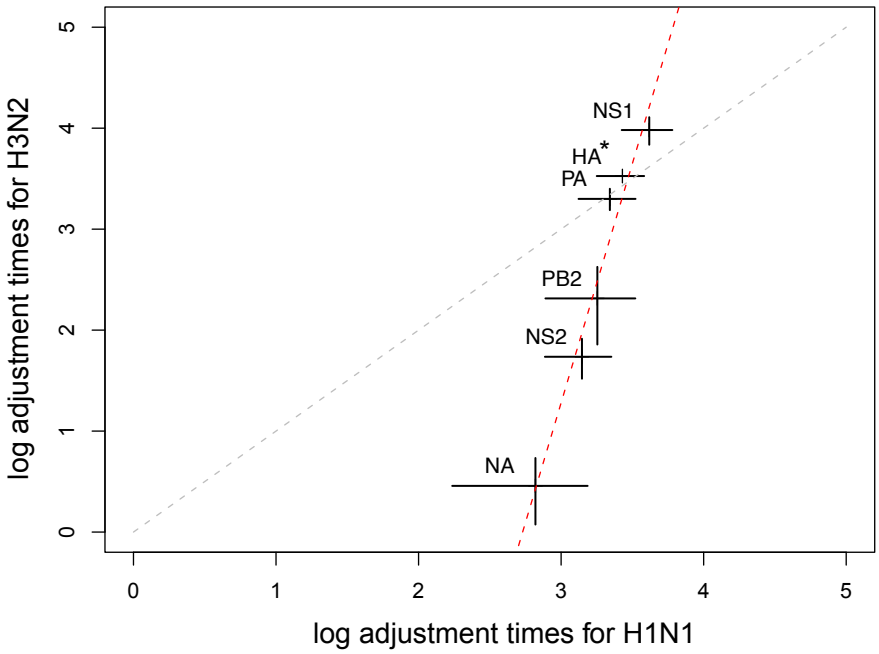

Figure 3.3: **GC3 adjustment times of H3N2** *vs.* **H1N1 viruses.** Results are shown on a $\log_e$-$\log_e$ scale (in years). The gray line represents the first bisector (line of equation $y = x$), while the red line represents the linear fit to the data. Bars: SEMs (95% Highest Posterior Densities, not shown, tend to be larger – see Fig. S27). $\star$: the HA value for H3N2 was tentatively derived using branches around the root node.

in terms of rates of evolution (Fig. 3.1 *vs.* 3.3). While NA is the fastest adjusting gene, HA appears to be the second slowest adjusting gene when information around the root node is used (Methods), ranking just after NS1. While the position of NS1 is consistent with a previous study [123], that of HA is in contrast to its high rate of evolution and the body of literature implicating HA in host preference, or to the idea that HA and NA need to be co-evolving as they both target the same sialic acids on the host cells [124]. Here however, genes are not ordered with respect to their importance in evading ongoing host immune responses or other form of adaptation, but with respect to how their GC3 content adjusts to that of their host. One potential explanation of the difference between evolution and adjustment rates is that highly expressed influenza genes adjust rapidly

since the virus highjacks the host translation machinery. However, while a number of studies have examined expression patterns of hosts genes [125], very little is known about expression patterns of viral genes during the course of an infection. Further research in viral transcriptomics is therefore warranted.

Second, some of the H1N1 genomes included in our alignment come from human viruses that reappeared in 1977 after a 20-year gap (*e.g.*, [112]). The presence of these genomes in our data could potentially bias downwards our estimates of adjustment times for H1N1 viruses, or at least increase the variance of these time estimates [126]. However, none of the H1N1 viruses that reappeared in 1977 underwent a change of host, so that these viruses were not included in our calculation of adjustment times. More crucially, removing these genomes (in gray in Fig. S1-S10) from the analyses estimating divergence times did not alter our estimates of t_{MRCA}, the age of the root (Fig. S27). Furthermore, previous work showed that the rate of evolution of these reintroduced sequences is similar to that of seasonal H1N1 sequences [112]. Altogether, our results are therefore robust to the presence of these re-emergent viruses.

Third, GC3 adjustment times also depend on the direction of host change (Fig. 3.4). This little-studied aspect outside of human transmission [114] reveals that across both H1N1 and H3N2 subtypes, adjustment of human viruses to avian hosts is the slowest, while adjustment of viruses coming from an avian host is very fast, with an average < 10 years (Fig. 3.4). On the other hand, subtypes show a difference in the GC3 adjustment speed of viruses coming from swine, with an average of 35 years for the H1N1 subtype *vs.* 5 years for H3N2. This difference between the adjustment dynamics of avian and swine viruses is somewhat unexpected, as swine is often considered to be the 'mixing vessel', harboring both types of sialic acids in its respiratory tracts and being therefore able to be infected by both avian and human viruses [127]. However, even if cross-

species transmission requires some adaptive process, we show next that GC3 adjustment is probably not adaptive *per se*.

3.6 GC3 adjustment reflects relaxed selective pressures

As viruses use the translational machinery of their host to translate their own mRNA, their codon usage and hence their GC3 content is expected to be under selective pressure to adapt to the pool of transfer RNA of their host [128]. To assess the role of selection during the GC3 content adjustment process, we tested for evidence of positive selection along the lineages starting from cross-species transmission and ending at the GC3 cluster change. Table 2.4 shows that no such evidence could be detected. This result could be due to (i) the inclusion of > 1 consecutive branches in the foreground lineages, (ii) selective forces acting on the background branches or to (iii) the non-distinction of the different directions of host change in this particular test. Current codon models allow us to have only one set of foreground branches [109], while with three hosts we would require six such sets, as done in the GC3 content analysis above (Fig. 3.4). It could also be possible to test for all possible combinations of foreground branches as recently proposed [129]. This procedure would circumvent the issue of using the same data twice, once to identify branches of interest and a second time to test for positive selection. However, while that approach would identify the branches along which positive selection can be detected [129], it would fail to test the specific hypothesis of presence of positive selection in the lineages between the cross-species transmission and the GC3 cluster change.

More critically, we find that shorter log adjustment times are significantly correlated with higher estimates of selection coefficients in the case of H3N2 ($\hat{\omega}$ under H_1; $P =$

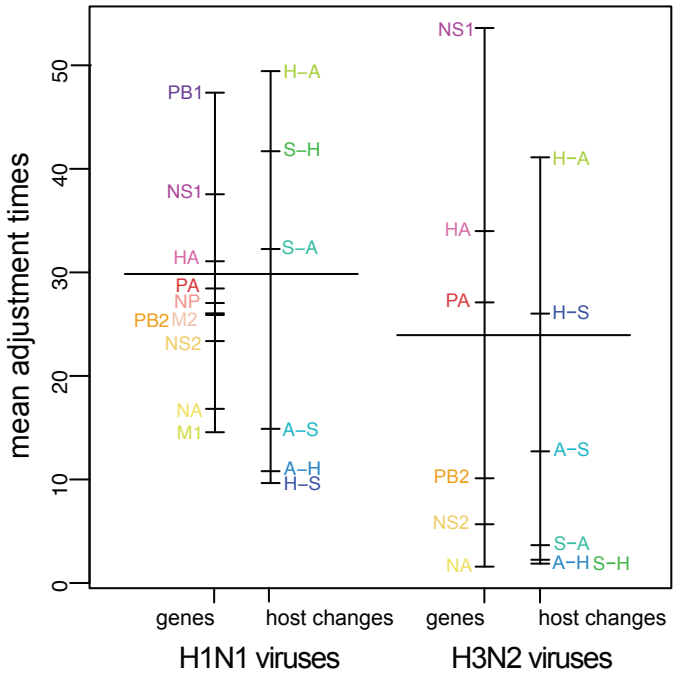

Figure 3.4: **Factor effects in the linear model (ANOVA) that was fitted to adjustment times (in years).** The directions of host change are avian-to-human (A-H), avian-to-swine (A-S), human-to-avian (H-A), human-to-swine (H-S), swine-to-avian (S-A) and swine-to-human (S-H). Adjustment times are in years. See text for details.

0.0066, $R^2 = 0.44$), but not in the case of H1N1 viruses ($P = 0.7995$). This result shows that relaxation of selective pressures plays a key role in the adjustment of H3N2 to a new host. While the lack of signal for H1N1 viruses might be due to ω rate ratios that are specific to the direction of host change, the population genetics of the two subtypes might also explain the difference. While both H1N1 and H3N2 viruses are expected to undergo frequent bottlenecks during their spread among hosts populations (and hence increase drift), the larger population sizes found in H3N2, the dominant subtype for > 40 years since the 1968 pandemic, are expected to facilitate the action of selection. Altogether, the differential incidence between the two subtypes could explain the stronger role of relaxation of selective pressures in H3N2 across the three hosts studied here.

The origin of the adjustment process can be revealed by considering the effective number of codons (ENC). In spite of most relationships between ENC and GC3 being significantly positive (Fig. S24), our data show no evidence for codon bias. Indeed, for the genes sampled here, ENC is never below the 35 threshold, which is usually taken as an indicator of strong codon bias [130] and ENC is almost always above 50 (Fig. S25-26). Altogether, our results suggest that GC3 adjustment is essentially driven by mutational bias in H1N1 and H3N2 viruses, with a larger role of relaxed selective pressures in H3N2 viruses. Future work should focus on the differential dynamics of H1N1 and H3N2 subtypes, potentially taking inspiration from the use of nonhomogeneous models as in [112], but developed at the codon level.

Chapter 4

Conclusions

4.1 Principal findings

We present here one of the most comprehensive analysis of the time-scale of influenza virus evolution, which is an original method to uncover dynamics of codon bias adjustment after cross-species transmission. In our study, we analyzed the whole genome of the influenza viruses responsible for high mortality and morbidity rate from 1900-2009.

With our phylogenetic analysis of whole genome of influenza viruses, we have found different phylogenic trajectories for genes of H1N1 and H3N2 viruses. Our phylogenetic analysis on H1N1 and H3N2 viruses highlights high number of reassortment within and between subtypes.

We showed here that the gene-specific rates of evolution is systematically larger for H3N2 than H1N1 viruses but not significantly ($P = 0.5245$). However the rate difference between the two subtypes is significant (at the 5% level) at least for three genes (HA, NA and NS2), which might be due to stronger selective pressure and dominancy of H3N2 viruses in human population for more than 40 years. We found a linear relationship between the gene-specific rates of evolution of H3N2 and H1N1 subtypes ($P = 6.8 \times 10^{-5}$;

$R^2 = 0.66$). One possible explanation is that the HA and NA which are the fast-evolving genes, appear at the surface of the viral particles and are directly involved in the immune escape of the virus, while the slow-evolving genes are involved in internal functions. NS2 also which is an internal fast-evolving protein involves in the immune escape of the virus with the host Fig. 3.1.

We detected that the GC3 contents of avian, human and swine viruses are significantly different ($F_{2,77179} = 862.81$, $P < 2.2 \times 10^{-16}$), with avian having the largest GC3 content, followed by swine. After cross-species transmission of the viruses, the GC3 content of viruses coming from different hosts tends to reflect that of heir host. GC3 contents of influenza viruses coming from specific hosts show the same trend as of their hosts.

We found a ($\log_e$-$\log_e$) linear relationship in terms of GC3 adjustment durations between the two subtypes as above for the rates of evolution. GC3 content of H3N2 viruses adjusts faster than in H1N1 viruses ($F_{4,5} = 67.65$, $P = 0.0012$) and the same ordering of genes exists for both subtypes ($P = 0.0006$; $R^2 = 0.96$).

Notably, the genes that are fast adjusting (e.g., NS2 and NA) are the ones involved in the final stages of the viral cycle. By fitting an ANOVA linear model we represented here mean adjustment times as a function of gene segment and direction of host changes for both subtypes. Results showed that these two factors have a highly significant effect ($P = 1.25 \times 10^{-6}$ and 5.74×10^{-10}).

Our novel approach for estimating the adjustment rate represented here have shown that the rank ordering of genes by their adjustment time differs from their ranking in terms of rates of evolution. For instance, HA and NA show the highest rate of evolution among other genes of influenza viruses, while NA is the fastest adjusting gene, HA appears to be the second slowest adjusting gene. One potential explanation of the difference between evolution and adjustment rates is that highly expressed influenza

genes adjust rapidly since the virus highjacks the host translation machinery.

We found that GC3 adjustment times depend on the direction of host change. For instance across both H1N1 and H3N2 subtypes, adjustment of human viruses to avian hosts is the slowest, while adjustment of viruses coming from an avian host is very fast. In addition, H1N1 and H3N2 subtypes show a difference in the GC3 adjustment speed of viruses coming from swine, however swine is often considered to be the 'mixing vessel', harboring both types of sialic acids in its respiratory tracts and being therefore able to be infected by both avian and human viruses.

Moreover, as a result of assessing the role of positive selection during the GC3 adjustment, we find that shorter log adjustment times are significantly correlated with higher estimates of selection coefficients in the case of H3N2 ($\hat{\omega}$ under H_1; $P = 0.0066$, $R^2 = 0.44$), but not in the case of H1N1 viruses ($P = 0.7995$). This result shows that relaxation of selective pressures plays a key role in the adjustment of H3N2 to a new host.

While the lack of signal for H1N1 viruses might be due to ω rate ratios that are specific to the direction of host change, the population genetics of the two subtypes might also explain the difference. While both H1N1 and H3N2 viruses are expected to undergo frequent bottlenecks during their spread among hosts populations (and hence increase drift), the larger population sizes found in H3N2, the dominant subtype for more than 40 years since the 1968 pandemic, are expected to facilitate the action of selection. Altogether, the differential incidence between the two subtypes could explain the stronger role of relaxation of selective pressures in H3N2 across the three hosts studied here.

4.2 Future directions

Systematic research on influenza over the past century has looked both forward and backward in time to identify viral evolution and transmission patterns. The more we learn about influenza viruses and their mortality and morbidity rate the more serious threat to human population they become. About a century after the fatal influenza pandemic of the 1918, still severe influenza pandemics and epidemics continue to emerge and treat human population.

Our novel approach presented here, to assess influenza's adjustment time opens an avenue to understand more about the dynamics of different subtypes of influenza viruses. Adjustment time patterns depicted here could be applied in more extensive studies on other viral hosts, subtypes, genes and more extensive global data sets, beyond North America.

Future work may focus on the differential dynamics of H1N1, H3N2 and other subtypes using different robust nonhomogeneous models at the codon level. More studies are required to unravel the main reason why we have a disconnect between the rate of evolution and the adjustment time across influenza virus genes. Adjustment process can be fast for some genes but not necessarily for the usual suspect, the HA gene. For example, the fastest evolving gene, which is HA, is not the fastest adjusting gene.

Here we found that dynamics (ordering) of genes are the same across subtypes, but its not clear why we have similar trend across subtypes. further work is required to unravel the cause of such a pattern, and to discern whether it is related to the viral population genetics or to the prevalence of different subtypes.

Also, our results here has potential surveillance and clinical implications in influenza virus drug and vaccine development. We speculate that the adjustment rate of different genes of influenza A viruses can be potentially considered as a proxy to find out the best

target for anti-viral drug and vaccines Fig. 3.1. Developing such drug and vaccines consequently might be the limiting step of the adjustment process to a new host. Moreover, our results can be applied in predictive study about influenza pandemics emergence, as dos Reis et al. (2011) suggested the degree of human adjustment of the virus plays an important role in host transmission to humans.

Finally, future work would benefit from incorporating regular surveillance, research, and international cooperation on preventing and overcoming influenza as a complicated large-scale challenge. Future in depth studies on influenza combining epidemiological, genomic, and antigenic data may enhance our understanding of early detection and control of new emerging strains.

4.3 Concluding statements

We showed here that studying cross-species transmissions of influenza A viruses that established themselves as stable lineages sheds some unsuspected light on the dynamics of two major subtypes. In particular, we demonstrated that both H1N1 and H3N2 subtypes have the same fast-adjusting genes in terms of GC3 content (Fig. 3.3), while H3N2 viruses adjust significantly faster (Fig. 3.3), in particular when coming from avian hosts (Fig. 3.4).

We also showed that two genes, NS2 and NA lead the pace of this adjustment process in both subtypes (Fig. 3.3). These genes play a key role in the final stages of the viral cycle in host cells (export of viral genome from nucleus and release of viral particles out of host cells, respectively), which consequently might be the limiting step of the adjustment process to a new host.

Although we did not attempt to validate the method on simulated data, extensions could consider using heterogeneous models [112]. Our results should also be validated

by analyzing other, more extensive, data sets, beyond North America, to confirm (i) the relationship between adjustment rates of H1N1 and H3N2 viruses and (ii) the disconnect between viral adjustability and evolutionary rate. Finally, our results highlight the importance of obtaining complete genome data through surveillance program in order to unravel the dynamics of influenza viruses, and not just from the standpoint of GC3 adjustment. We argue that only such complete genome information will help us understand how emerging pathogens acquire the ability to be efficiently transmitted within their new host [131]. The most likely answer may not lie in the identification of signature amino acid sites, but rather in the determination of epistatic interaction of sites within [132] and among segments [111].

Bibliography

[1] Watanabe Y, Ibrahim MS, Suzuki Y, Ikuta K: **The changing nature of avian influenza A virus (H5N1)**. *Trends in microbiology* 2012, **20**:11–20.

[2] McHardy AC, Adams B: **The role of genomics in tracking the evolution of influenza A virus**. *PLoS pathogens* 2009, **5**(10):e1000566.

[3] Coloma R, Valpuesta JM, Arranz R, Carrascosa JL, Ortín J, Martín-Benito J: **The structure of a biologically active influenza virus ribonucleoprotein complex**. *PLoS pathogens* 2009, **5**(6):e1000491.

[4] Nicholson K: **Human influenza**. *Textbook of influenza. London: Blackwell Scientific Publications* 1998, :219–66.

[5] Van-Tam J, Sellwood C: *Introduction to pandemic influenza*. Modular texts, Wallingford, Oxfordshire: CAB International 2010.

[6] Compans RW: *Vaccines for pandemic influenza*. New York: Springer 2009.

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

[8] Reid AH, Taubenberger JK: **The origin of the 1918 pandemic influenza virus: a continuing enigma**. *J Gen Virol* 2003, **84**(Pt 9):2285–92.

[9] Morens DM, Taubenberger JK, Harvey HA, Memoli MJ: **The 1918 influenza pandemic: lessons for 2009 and the future**. *Critical care medicine* 2010, **38**(4 Suppl):e10.

[10] Enemark C: **Is pandemic flu a security threat?** *Survival* 2009, **51**:191–214.

[11] Shrestha SS, Swerdlow DL, Borse RH, Prabhu VS, Finelli L, Atkins CY, Owusu-Edusei K, Bell B, Mead PS, Biggerstaff M, et al.: **Estimating the burden of 2009 pandemic influenza A (H1N1) in the United States (April 2009–April 2010)**. *Clinical Infectious Diseases* 2011, **52**(suppl 1):S75–S82.

[12] Dawood FS, Iuliano AD, Reed C, Meltzer MI, Shay DK, Cheng PY, Bandaranayake D, Breiman RF, Brooks WA, Buchy P, et al.: **Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study**. *The Lancet infectious diseases* 2012, **12**(9):687–695.

[13] Suzuki Y, Ito T, Suzuki T, Holland RE, Chambers TM, Kiso M, Ishida H, Kawaoka Y: **Sialic acid species as a determinant of the host range of influenza A viruses**. *Journal of virology* 2000, **74**(24):11825–11831.

[14] Itzstein Mv: *Influenza virus sialidase: a drug discovery target*. Milestones in drug therapy, Basel: Springer 2012.

[15] Scholtissek C: **Molecular evolution of influenza viruses**. *Virus Genes* 1995, **11**(2-3):209–15.

[16] Belshe RB: **The origins of pandemic influenza–lessons from the 1918 virus**. *N Engl J Med* 2005, **353**(21):2209–11.

[17] Taubenberger JK, Reid AH, Lourens RM, Wang R, Jin G, Fanning TG: **Characterization of the 1918 influenza virus polymerase genes**. *Nature* 2005, **437**(7060):889–93.

[18] Matrosovich M, Tuzikov A, Bovin N, Gambaryan A, Klimov A, Castrucci MR, Donatelli I, Kawaoka Y: **Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals**. *J Virol* 2000, **74**(18):8502–12.

[19] Stevens J, Blixt O, Tumpey TM, Taubenberger JK, Paulson JC, Wilson IA: **Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus**. *Science* 2006, **312**(5772):404–10.

[20] Tumpey TM, Maines TR, Van Hoeven N, Glaser L, Solórzano A, Pappas C, Cox NJ, Swayne DE, Palese P, Katz JM, García-Sastre A: **A two-amino acid change in the hemagglutinin of the 1918 influenza virus abolishes transmission**. *Science* 2007, **315**(5812):655–9.

[21] Finkelstein DB, Mukatira S, Mehta PK, Obenauer JC, Su X, Webster RG, Naeve CW: **Persistent host markers in pandemic and H5N1 influenza viruses**. *J Virol* 2007, **81**(19):10292–9.

[22] Imai M, Watanabe T, Hatta M, Das SC, Ozawa M, Shinya K, Zhong G, Hanson A, Katsura H, Watanabe S, et al.: **Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets**. *Nature* 2012, **486**(7403):420–428.

[23] Osterhaus A, Rimmelzwaan G, Martina B, Bestebroer T, Fouchier R: **Influenza B virus in seals**. *Science* 2000, **288**(5468):1051–1053.

[24] Hay AJ, Gregory V, Douglas AR, Lin YP: **The evolution of human influenza viruses**. *Philos Trans R Soc Lond B Biol Sci* 2001, **356**(1416):1861–70.

[25] Yamashita M, Krystal M, Fitch WM, Palese P: **Influenza B virus evolution: co-circulating lineages and comparison of evolutionary pattern with those of influenza A and C viruses**. *Virology* 1988, **163**:112–122.

[26] Yuanji G, Fengen J, Ping W, Min W, Jiming Z: **Isolation of influenza C virus from pigs and experimental infection of pigs with influenza C virus**. *Journal of General Virology* 1983, **64**:177–182.

[27] Tong S, Zhu X, Li Y, Shi M, Zhang J, Bourgeois M, Yang H, Chen X, Recuenco S, Gomez J, Chen LM, Johnson A, Tao Y, Dreyfus C, Yu W, McBride R, Carney PJ, Gilbert AT, Chang J, Guo Z, Davis CT, Paulson JC, Stevens J, Rupprecht CE, Holmes EC, Wilson IA, Donis RO: **New world bats harbor diverse influenza A viruses**. *PLoS Pathog* 2013, **9**(10):e1003657.

[28] Ito T, Kawaoka Y: **Host-range barrier of influenza A viruses**. *Veterinary microbiology* 2000, **74**:71–75.

[29] Suzuki Y, Nei M: **Origin and evolution of influenza virus hemagglutinin genes**. *Molecular biology and evolution* 2002, **19**(4):501–509.

[30] Palese P: **Influenza: old and new threats**. *Nature medicine* 2004, **10**:S82–S87.

[31] Jagger B, Wise H, Kash J, Walters KA, Wills N, Xiao YL, Dunfee R, Schwartzman L, Ozinsky A, Bell G, et al.: **An overlapping protein-coding region in influenza A virus segment 3 modulates the host response**. *Science* 2012, **337**(6091):199–204.

[32] Wise HM, Hutchinson EC, Jagger BW, Stuart AD, Kang ZH, Robb N, Schwartzman LM, Kash JC, Fodor E, Firth AE, et al.: **Identification of a novel splice variant form of the influenza a virus m2 ion channel with an antigenically distinct ectodomain**. *PLoS pathogens* 2012, **8**(11):e1002998.

[33] Herfst S, Schrauwen EJ, Linster M, Chutinimitkul S, de Wit E, Munster VJ, Sorrell EM, Bestebroer TM, Burke DF, Smith DJ, et al.: **Airborne transmission of influenza A/H5N1 virus between ferrets**. *Science* 2012, **336**(6088):1534–1541.

[34] Shaw M, Palese P: **Orthomyxoviruses: molecular biology**. *Encyclopedia of virology* 2008, **3**:483–489.

[35] Krug R, Lamb R: **Orthomyxoviridae: the viruses and their replication**. *Fields Virology* 2001, :1487–1503.

[36] Poole E, Elton D, Medcalf L, Digard P: **Functional domains of the influenza A virus PB2 protein: identification of NP- and PB1-binding sites**. *Virology* 2004, **321**:120–33.

[37] Ng AKL, Chan WH, Choi ST, Lam MKH, Lau KF, Chan PKS, Au SWN, Fodor E, Shaw PC: **Influenza polymerase activity correlates with the strength of interaction between nucleoprotein and PB2 through the host-specific residue K/E627**. *PLoS One* 2012, **7**(5):e36415.

[38] Wise HM, Foeglein A, Sun J, Dalton RM, Patel S, Howard W, Anderson EC, Barclay WS, Digard P: **A complicated message: Identification of a novel PB1-related protein translated from influenza A virus segment 2 mRNA**. *J Virol* 2009, **83**(16):8021–31.

[39] Smith AM, Adler FR, McAuley JL, Gutenkunst RN, Ribeiro RM, McCullers JA, Perelson AS: **Effect of 1918 PB1-F2 expression on influenza A virus infection kinetics**. *PLoS computational biology* 2011, **7**(2):e1001081.

[40] Eckert DM, Kim PS: **Mechanisms of viral membrane fusion and its inhibition**. *Annual review of biochemistry* 2001, **70**:777–810.

[41] Watanabe Y, Ibrahim MS, Ellakany HF, Kawashita N, Mizuike R, Hiramatsu H, Sriwilaijaroen N, Takagi T, Suzuki Y, Ikuta K: **Acquisition of human-type receptor binding specificity by new H5N1 influenza virus sublineages during their emergence in birds in Egypt**. *PLoS pathogens* 2011, **7**(5):e1002068.

[42] Medcalf L, Poole E, Elton D, Digard P: **Temperature-sensitive lesions in two influenza A viruses defective for replicative transcription disrupt RNA binding by the nucleoprotein**. *Journal of virology* 1999, **73**(9):7349–7356.

[43] Portela A, Digard P: **The influenza virus nucleoprotein: a multifunctional RNA-binding protein pivotal to virus replication**. *Journal of General Virology* 2002, **83**(4):723–734.

[44] Huang IC, Li W, Sui J, Marasco W, Choe H, Farzan M: **Influenza A virus neuraminidase limits viral superinfection**. *Journal of virology* 2008, **82**(10):4834–4843.

[45] Ruigrok RW, Barge A, Durrer P, Brunner J, Ma K, Whittaker GR: **Membrane interaction of influenza virus M1 protein**. *Virology* 2000, **267**(2):289–298.

[46] Zhang J, Lamb RA: **Characterization of the membrane association of the influenza virus matrix protein in living cells**. *Virology* 1996, **225**(2):255–266.

[47] Gómez-Puertas P, Albo C, Pérez-Pastrana E, Vivo A, Portela A: **Influenza virus matrix protein is the major driving force in virus budding**. *Journal of virology* 2000, **74**(24):11538–11547.

[48] Lamb RA, Zebedee SL, Richardson CD: **Influenza virus M2 protein is an integral membrane protein expressed on the infected-cell surface**. *Cell* 1985, **40**(3):627–633.

[49] Ichinohe T, Pang IK, Iwasaki A: **Influenza virus activates inflammasomes via its intracellular M2 ion channel**. *Nature immunology* 2010, **11**(5):404–410.

[50] LU Y, WAMBACH M, KATZE MG, KRUG RM: **Binding of the influenza virus NS1 protein to double-stranded RNA inhibits the activation of the protein kinase that phosphorylates the eIF-2 translation initiation factor**. *Virology* 1995, **214**:222–228.

[51] Paterson D, Fodor E: **Emerging roles for the influenza A virus nuclear export protein (NEP)**. *PLoS Pathog* 2012, **8**(12):e1003019.

[52] Lamb RA, Choppin PW, Chanock RM, Lai CJ: **Mapping of the two overlapping genes for polypeptides NS1 and NS2 on RNA segment 8 of influenza virus genome**. *Proc Natl Acad Sci U S A* 1980, **77**(4):1857–61.

[53] Ward AC, Castelli LA, Lucantoni AC, White JF, Azad AA, Macreadie IG: **Expression and analysis of the NS2 protein of influenza A virus**. *Arch Virol* 1995, **140**(11):2067–73.

[54] O'Neill RE, Talon J, Palese P: **The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins**. *EMBO J* 1998, **17**:288–96.

[55] O'Neill RE, Talon J, Palese P: **The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins**. *The EMBO journal* 1998, **17**:288–296.

[56] Samji T: **Influenza A: understanding the viral life cycle**. *The Yale journal of biology and medicine* 2009, **82**(4):153.

[57] Gerloff NA, Jones J, Simpson N, Balish A, Elbadry MA, Baghat V, Rusev I, de Mattos CC, de Mattos CA, Zonkle LEA, Kis Z, Davis CT, Yingst S, Cornelius C, Soliman A, Mohareb E, Klimov A, Donis RO: **A high diversity of Eurasian lineage low pathogenicity avian influenza A viruses circulate among wild birds sampled in Egypt**. *PLoS One* 2013, **8**(7):e68522.

[58] Gorman O, Bean W, Kawaoka Y, Donatelli I, Guo Y, Webster R: **Evolution of influenza A virus nucleoprotein genes: implications for the origins of H1N1 human and classical swine viruses.** *Journal of virology* 1991, **65**(7):3704–3714.

[59] Taubenberger JK, Morens DM: **Influenza: the once and future pandemic**. *Public Health Rep* 2010, **125 Suppl 3**:16–26.

[60] Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y: **Evolution and ecology of influenza A viruses**. *Microbiological reviews* 1992, **56**:152–179.

[61] Chen R, Holmes EC: **Avian influenza virus exhibits rapid evolutionary dynamics**. *Mol Biol Evol* 2006, **23**(12):2336–41.

[62] Holmes EC: *The evolution and emergence of RNA viruses*. Oxford series in ecology and evolution, Oxford: Oxford University Press 2009.

[63] Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, Holmes EC: **Unifying the epidemiological and evolutionary dynamics of pathogens**. *Science* 2004, **303**(5656):327–332.

[64] Cox N, Subbarao K: **Global epidemiology of influenza: past and present**. *Annual review of medicine* 2000, **51**:407–421.

[65] Nelson MI, Holmes EC: **The evolution of epidemic influenza**. *Nature reviews genetics* 2007, **8**(3):196–205.

[66] Nelson MI, Viboud C, Simonsen L, Bennett RT, Griesemer SB, George KS, Taylor J, Spiro DJ, Sengamalay NA, Ghedin E, et al.: **Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918**. *PLoS pathogens* 2008, **4**(2):e1000012.

[67] Lofgren E, Fefferman NH, Naumov YN, Gorski J, Naumova EN: **Influenza seasonality: underlying causes and modeling theories**. *J Virol* 2007, **81**(11):5429–36.

[68] Webster RG, Sharp GB, Claas EC: **Interspecies transmission of influenza viruses**. *American journal of respiratory and critical care medicine* 1995, **152**(4):S25.

[69] Scholtissek C, Rohde Wv, Von Hoyningen V, Rott R: **On the origin of the human influenza virus subtypes H2N2 and H3N2**. *Virology* 1978, **87**:13–20.

[70] Kawaoka Y, Krauss S, Webster RG: **Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics.** *Journal of Virology* 1989, **63**(11):4603–4608.

[71] Reid AH, Taubenberger JK: **The origin of the 1918 pandemic influenza virus: a continuing enigma**. *Journal of General Virology* 2003, **84**(9):2285–2292.

[72] Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, Sessions WM, Xu X, Skepner E, Deyde V, et al.: **Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans**. *science* 2009, **325**(5937):197–201.

[73] Holder M, Lewis PO: **Phylogeny estimation: traditional and Bayesian approaches**. *Nature reviews genetics* 2003, **4**(4):275–284.

[74] Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Molecular biology and evolution* 1987, **4**(4):406–425.

[75] Fitch WM: **Toward defining the course of evolution: minimum change for a specific tree topology**. *Systematic Biology* 1971, **20**(4):406–416.

[76] Fienberg SE: **When did Bayesian inference become" Bayesian"?** *Bayesian analysis* 2006, **1**:1–40.

[77] Dimmock NJ, Easton AJ, Leppard K: *Introduction to modern virology*. Malden, MA: Blackwell Pub., 6th ed edition 2007, [http://www.loc.gov/catdir/toc/ecip0610/2006009426.html].

[78] Wong EHM, Smith DK, Rabadan R, Peiris M, Poon LLM: **Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus**. *BMC Evol Biol* 2010, **10**:253.

[79] Cardinale DJ, Duffy S: **Single-stranded genomic architecture constrains optimal codon usage**. *Bacteriophage* 2011, **1**(4):219–224.

[80] Lobo FP, Mota BEF, Pena SDJ, Azevedo V, Macedo AM, Tauch A, Machado CR, Franco GR: **Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts**. *PLoS One* 2009, **4**(7):e6282.

[81] Fancher KC, Hu W: **Codon bias of influenza a viruses and their hosts**. *American Journal of Molecular Biology* 2011, **1**:174–182.

[82] Kober KM, Pogson GH: **Genome-wide patterns of codon bias are shaped by natural selection in the purple sea urchin, Strongylocentrotus purpuratus**. *G3 (Bethesda)* 2013, **3**(7):1069–83.

[83] Brower-Sinning R, Carter DM, Crevar CJ, Ghedin E, Ross TM, Benos PV: **The role of RNA folding free energy in the evolution of the polymerase genes of the influenza A virus**. *Genome Biol* 2009, **10**(2):R18.

[84] Anhlan D, Grundmann N, Makalowski W, Ludwig S, Scholtissek C: **Origin of the 1918 pandemic H1N1 influenza A virus as studied by codon usage patterns and phylogenetic analysis**. *RNA* 2011, **17**:64–73.

[85] Rahnama L, Aris-Brosou S: **Phylodynamics of the emergence of influenza viruses after cross-species transmission**. *PLoS One* 2013, **8**(12):e82486.

[86] Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D: **The influenza virus resource at the National Center for Biotechnology Information**. *J Virol* 2008, **82**(2):596–601.

[87] Bastien N, Antonishyn NA, Brandt K, Wong CE, Chokani K, Vegh N, Horsman GB, Tyler S, Graham MR, Plummer FA, Levett PN, Li Y: **Human infection with a triple-reassortant swine influenza A(H1N1) virus containing the**

hemagglutinin and neuraminidase genes of seasonal influenza virus. *J Infect Dis* 2010, **201**(8):1178–82.

[88] Aris-Brosou S: **A simple measure of the dynamics of segmented genomes: An application to influenza**. *Lecture Notes in Computer Science* 2010, **6398 LNBI**:149–160.

[89] Abdussamad J, Aris-Brosou S: **The nonadaptive nature of the H1N1 2009 Swine Flu pandemic contrasts with the adaptive facilitation of transmission to a new host**. *BMC Evol Biol* 2011, **11**:6.

[90] Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792–7.

[91] Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W609–12.

[92] Swofford D: *PAUP Phylogenetic Analysis Using Parsimony (Version 4), Sinauer, Sunderland, MA.* 2003.

[93] Schloss PD, Handelsman J: **Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness**. *Appl Environ Microbiol* 2005, **71**(3):1501–6.

[94] Taubenberger JK, Reid AH, Krafft AE, Bijwaard KE, Fanning TG: **Initial genetic characterization of the 1918 "Spanish" influenza virus**. *Science* 1997, **275**(5307):1793–6.

[95] Posada D: **Selection of models of DNA evolution with jModelTest**. *Methods Mol Biol* 2009, **537**:93–112.

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

[97] Rambaut A: **Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies**. *Bioinformatics* 2000, **16**(4):395–9.

[98] Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees**. *BMC Evol Biol* 2007, **7**:214.

[99] Drummond AJ, Ho SYW, Phillips MJ, Rambaut A: **Relaxed phylogenetics and dating with confidence**. *PLoS Biol* 2006, **4**(5):e88.

[100] Drummond AJ, Rambaut A, Shapiro B, Pybus OG: **Bayesian coalescent inference of past population dynamics from molecular sequences**. *Mol Biol Evol* 2005, **22**(5):1185–92.

[101] Huelsenbeck JP, Bollback JP, Levine AM: **Inferring the root of a phylogenetic tree**. *Syst Biol* 2002, **51**:32–43.

[102] Pagel M: **Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters**. *Proceedings: Biological Sciences* 1994, **255**(1342):37–45, [http://www.jstor.org/stable/49836].

[103] Schluter D, Price T, Mooers AØ, Ludwig D: **Likelihood of ancestor states in adaptive radiation**. *Evolution* 1997, :1699–1711.

[104] Paradis E: *Analysis of phylogenetics and evolution with R*. New York: Springer 2006.

[105] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2011.

[106] McInerney JO: **GCUA: general codon usage analysis**. *Bioinformatics* 1998, **14**(4):372–3.

[107] Aris-Brosou S: **Dating phylogenies with hybrid local molecular clocks**. *PLoS One* 2007, **2**(9):e879.

[108] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Spudich G, Tang YA, Trevanion S, Vandrovcova J, Vilella AJ, White S, Wilder SP, Zadissa A, Zamora J, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Herrero J, Hubbard TJP, Parker A, Proctor G, Vogel J, Searle SMJ: **Ensembl 2011**. *Nucleic Acids Res* 2011, **39**(Database issue):D800–6.

[109] Zhang J, Nielsen R, Yang Z: **Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level**. *Mol Biol Evol* 2005, **22**(12):2472–9.

[110] Yang Z, Wong WSW, Nielsen R: **Bayes empirical Bayes inference of amino acid sites under positive selection**. *Mol Biol Evol* 2005, **22**(4):1107–18.

[111] Yohai VJ: **High breakdown-point and high efficiency robust estimates for regression**. *The Annals of Statistics* 1987, :642–656.

[112] dos Reis M, Hay AJ, Goldstein RA: **Using non-homogeneous models of nucleotide substitution to identify host shift events: application to the origin of the 1918 'Spanish' influenza pandemic virus**. *J Mol Evol* 2009, **69**(4):333–45.

[113] Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghwani J, Bhatt S, Peiris JSM, Guan Y, Rambaut A: **Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic**. *Nature* 2009, **459**(7250):1122–5.

[114] Vijaykrishna D, Smith GJD, Pybus OG, Zhu H, Bhatt S, Poon LLM, Riley S, Bahl J, Ma SK, Cheung CL, Perera RAPM, Chen H, Shortridge KF, Webby RJ, Webster RG, Guan Y, Peiris JSM: **Long-term evolution and transmission dynamics of swine influenza A virus**. *Nature* 2011, **473**(7348):519–22.

[115] Olsen B, Munster VJ, Wallensten A, Waldenström J, Osterhaus ADME, Fouchier RAM: **Global patterns of influenza A virus in wild birds**. *Science* 2006, **312**(5772):384–8.

[116] Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC: **The genomic and epidemiological dynamics of human influenza A virus**. *Nature* 2008, **453**(7195):615–9.

[117] Ferguson NM, Galvani AP, Bush RM: **Ecological and immunological determinants of influenza evolution**. *Nature* 2003, **422**(6930):428–33.

[118] Zuckerkandl E, Pauling L: **Evolutionary divergence and convergence in proteins**. In *Evolving Genes and Proteins*. Edited by Bryson V, Vogel HJ, Academic Press 1965:97–166.

[119] Neumann G, Hughes MT, Kawaoka Y: **Influenza A virus NS2 protein mediates vRNP nuclear export through NES-independent interaction with hCRM1**. *EMBO J* 2000, **19**(24):6751–8.

[120] Nguyen KL, llano M, Akari H, Miyagi E, Poeschla EM, Strebel K, Bour S: **Codon optimization of the HIV-1 vpu and vif genes stabilizes their mRNA and allows for highly efficient Rev-independent expression**. *Virology* 2004, **319**(2):163–75.

[121] Lucks JB, Nelson DR, Kudla GR, Plotkin JB: **Genome landscapes and bacteriophage codon usage**. *PLoS Comput Biol* 2008, **4**(2):e1000001.

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

[123] dos Reis M, Tamuri AU, Hay AJ, Goldstein RA: **Charting the host adaptation of influenza viruses**. *Mol Biol Evol* 2011, **28**(6):1755–67.

[124] Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, Calisher CH, Laughlin CA, Saif LJ, Daszak P: **Cross-species virus transmission and the emergence of new epidemic diseases**. *Microbiol Mol Biol Rev* 2008, **72**(3):457–70.

[125] Kawada Ji, Kimura H, Kamachi Y, Nishikawa K, Taniguchi M, Nagaoka K, Kurahashi H, Kojima S, Morishima T: **Analysis of gene-expression profiles by oligonucleotide microarray in children with influenza**. *Journal of general virology* 2006, **87**(6):1677–1683.

[126] Wertheim JO: **The re-emergence of H1N1 influenza virus in 1977: a cautionary tale for estimating divergence times using biologically unrealistic sampling dates**. *PLoS One* 2010, **5**(6):e11184.

[127] Scholtissek C, Bürger H, Kistner O, Shortridge KF: **The nucleoprotein as a possible major factor in determining host specificity of influenza H3N2 viruses**. *Virology* 1985, **147**(2):287–94.

[128] Sharp PM, Li WH: **The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications**. *Nucleic acids research* 1987, **15**(3):1281–1295.

[129] Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K: **A random effects branch-site model for detecting episodic diversifying selection**. *Mol Biol Evol* 2011.

[130] Wright F: **The 'effective number of codons' used in a gene**. *Gene* 1990, **87**:23–9.

[131] Wolfe ND, Dunavan CP, Diamond J: **Origins of major human infectious diseases**. *Nature* 2007, **447**(7142):279–83.

[132] Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB: **Prevalence of epistasis in the evolution of influenza A surface proteins**. *PLoS Genet* 2011, **7**(2):e1001301.

Appendices

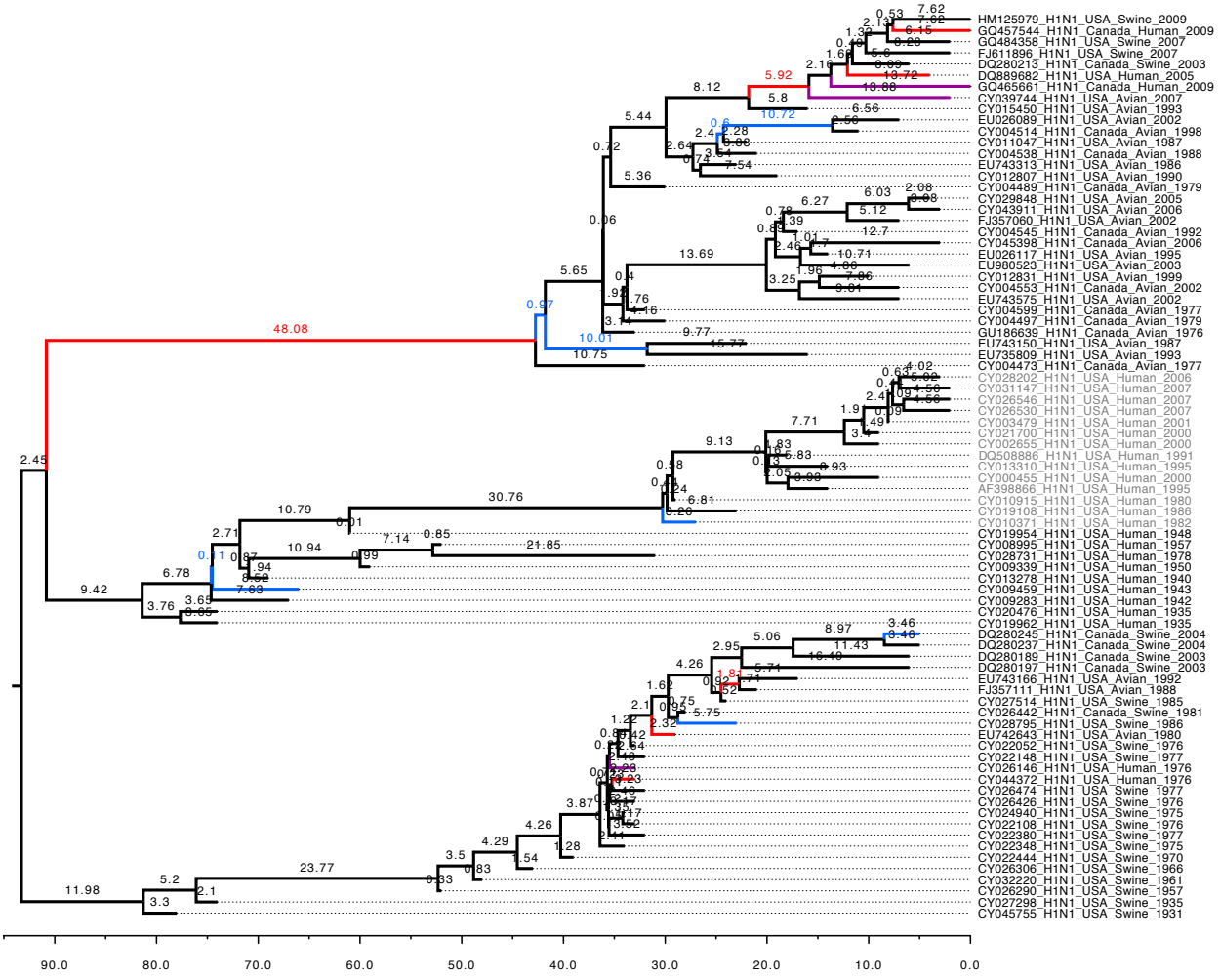

Figure S1: **Timed tree for subtype H1N1 gene PB2.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray.

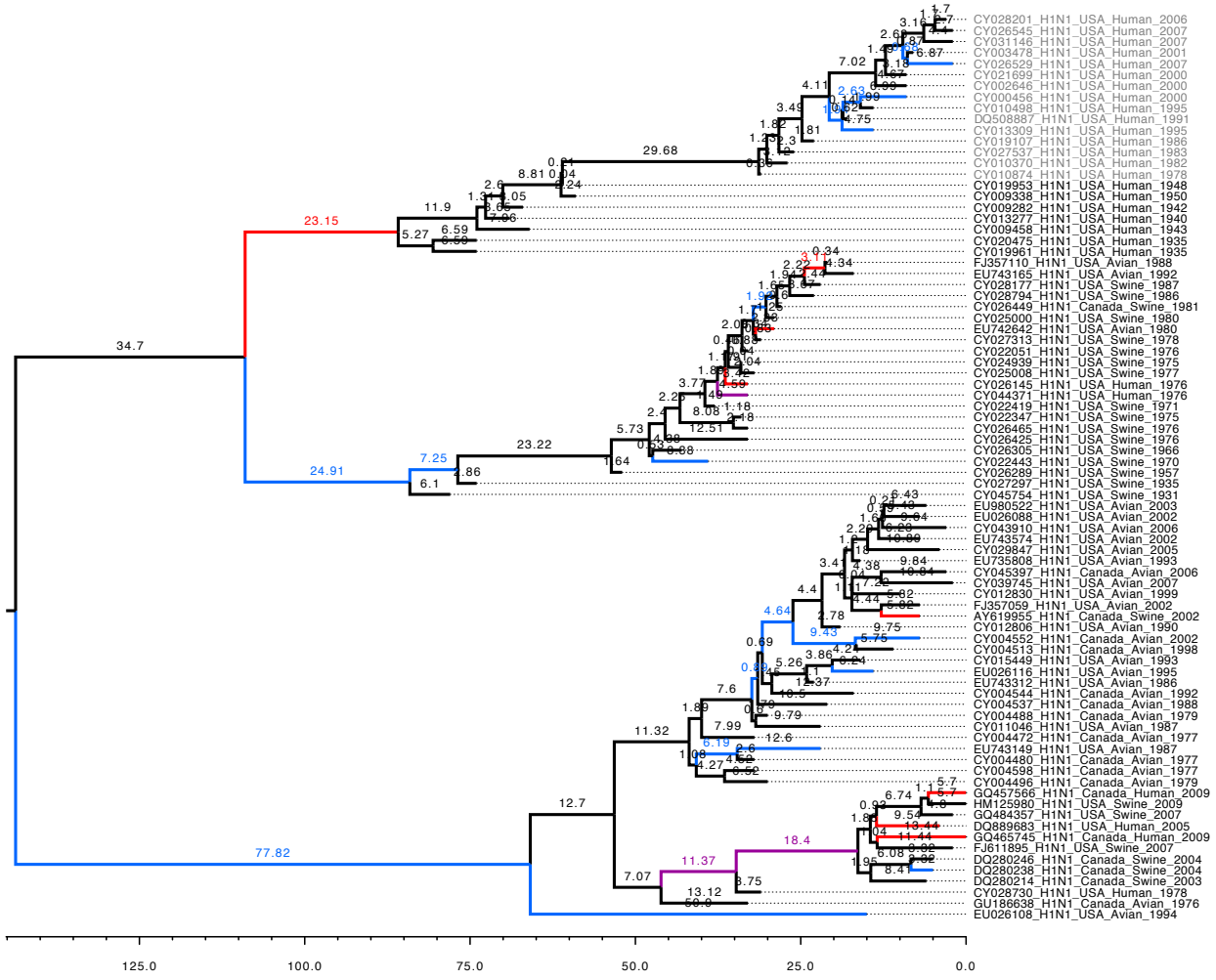

Figure S2: **Timed tree for subtype H1N1 gene PB1.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray.

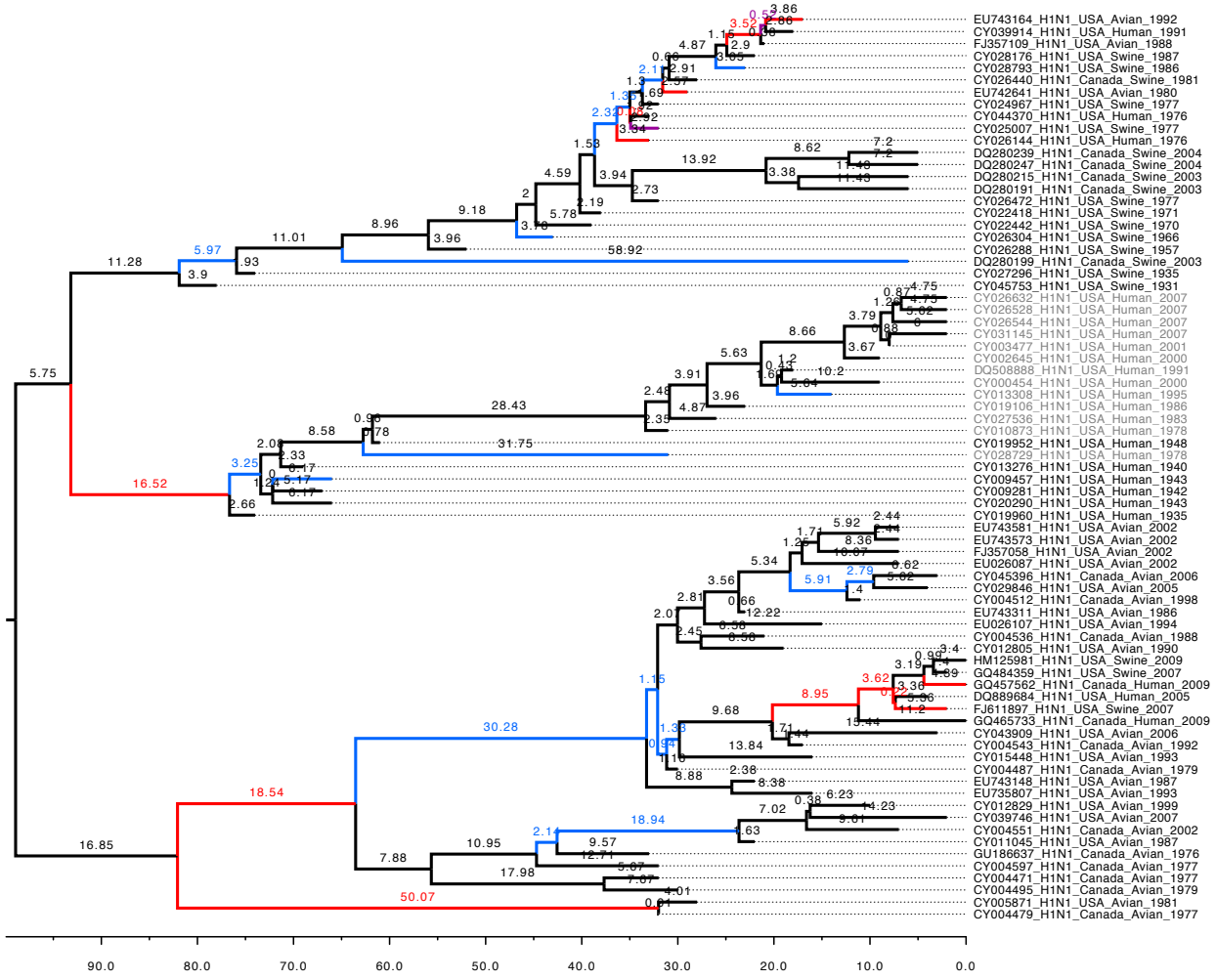

Figure S3: **Timed tree for subtype H1N1 gene PA.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray.

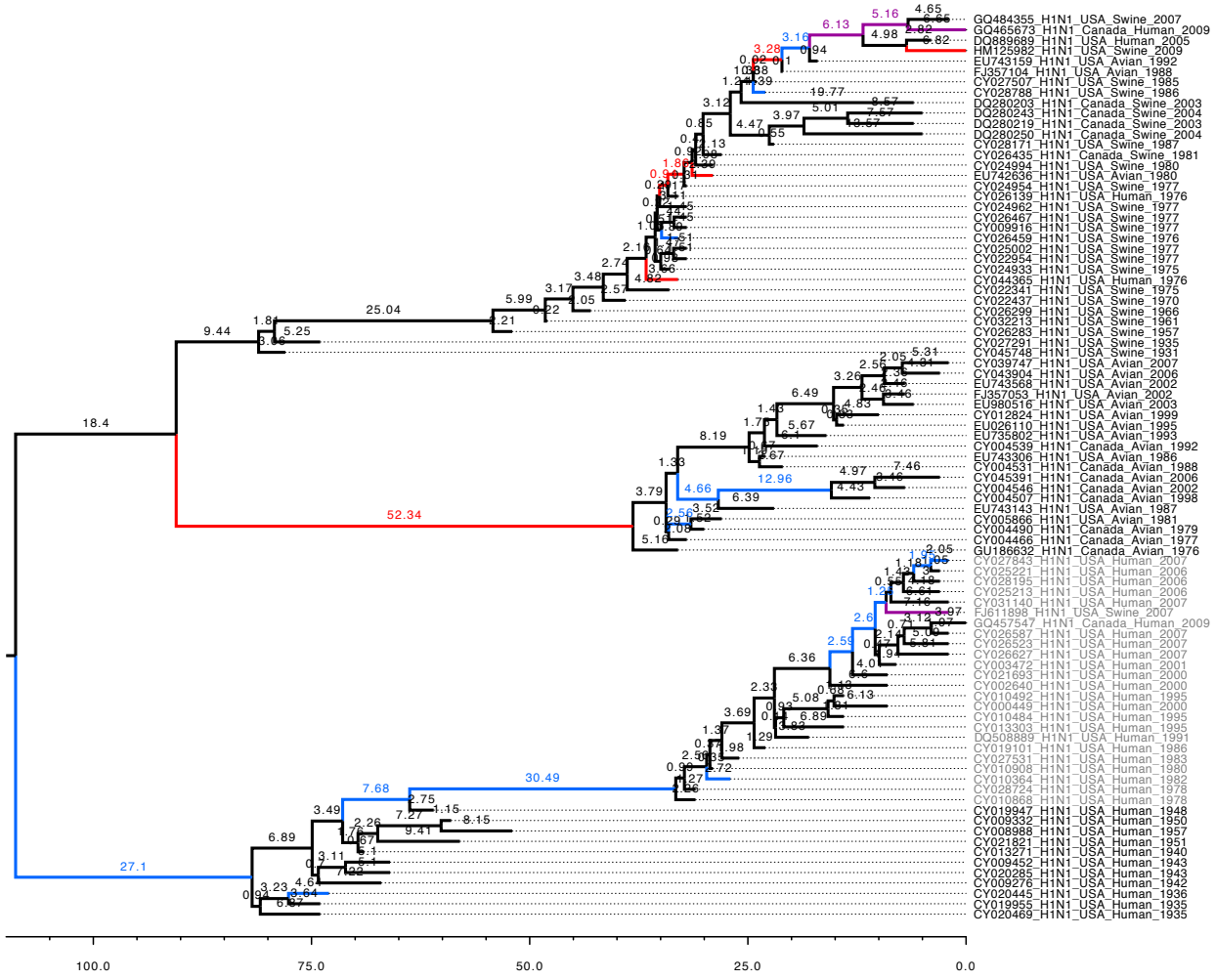

Figure S4: **Timed tree for subtype H1N1 gene HA.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray.

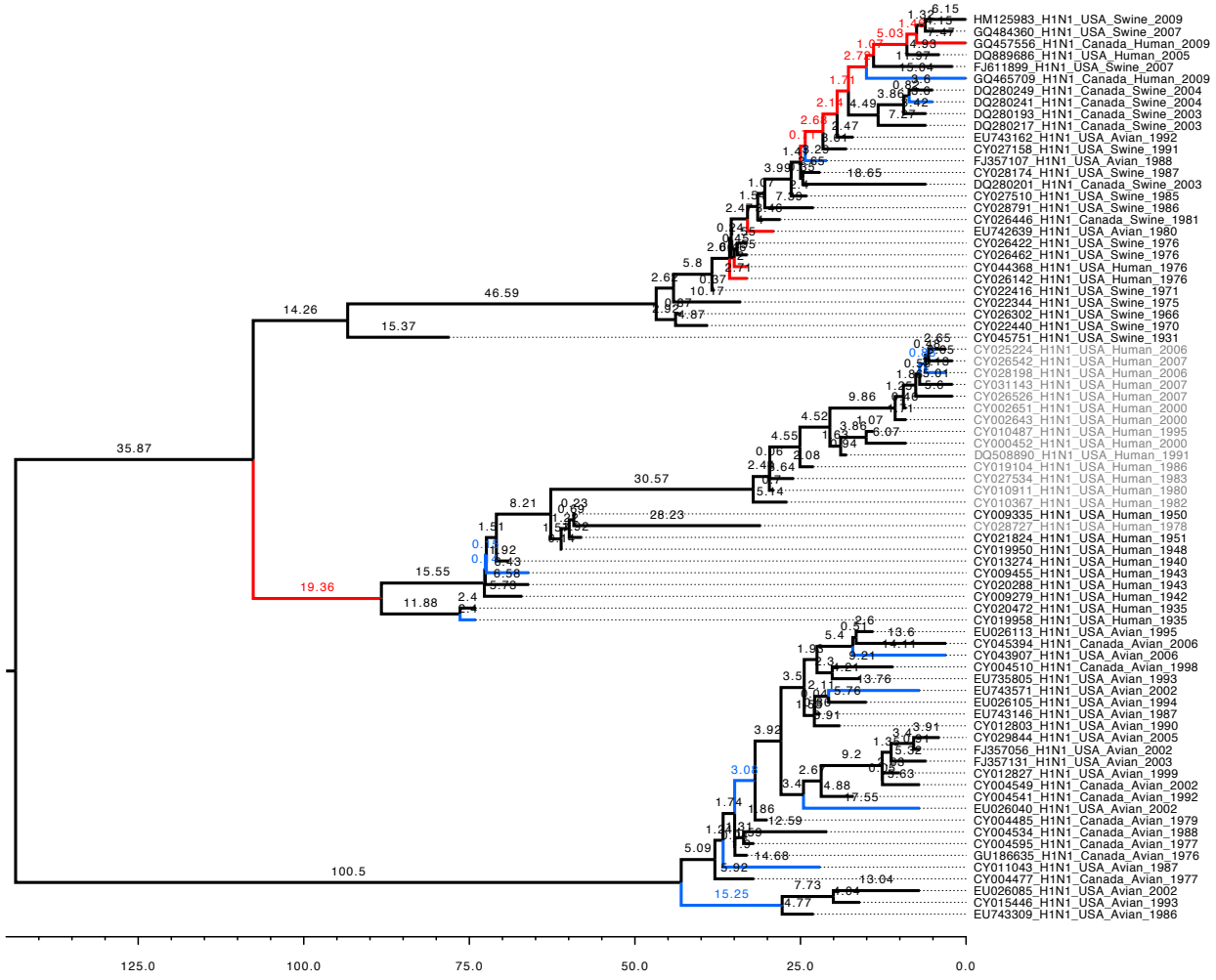

Figure S5: **Timed tree for subtype H1N1 gene NP.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray.

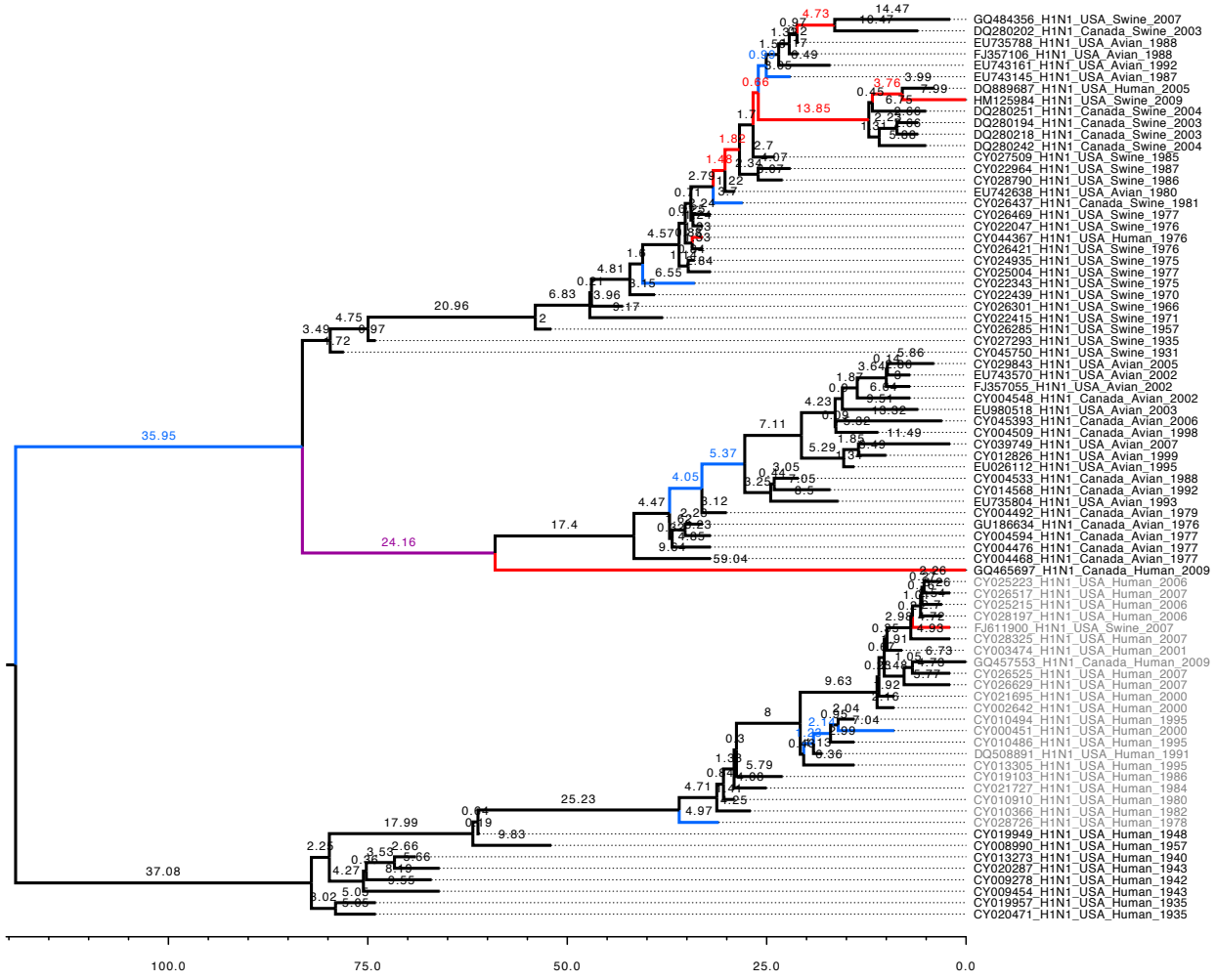

Figure S6: **Timed tree for subtype H1N1 gene NA.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray.

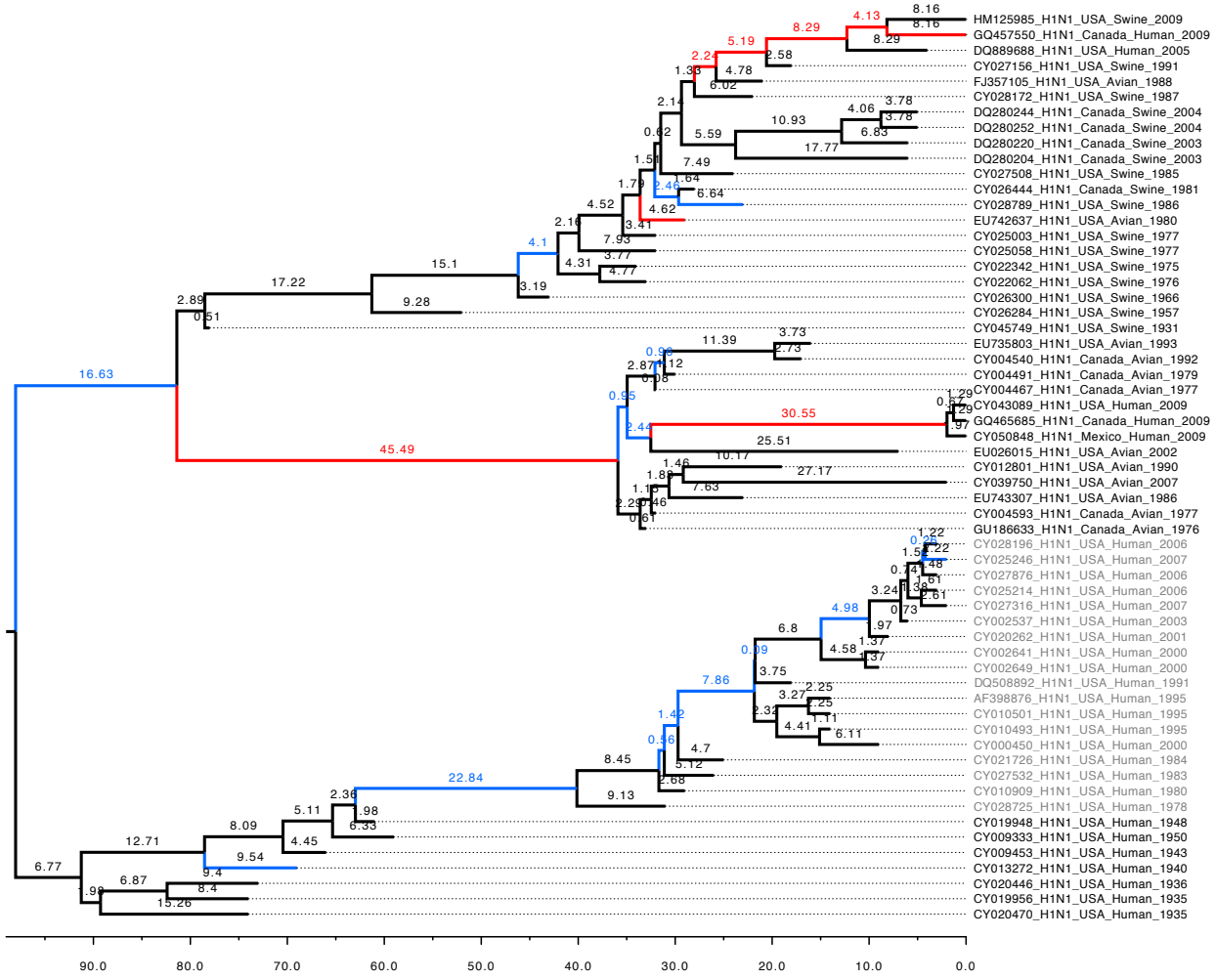

Figure S7: **Timed tree for subtype H1N1 gene M2.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray.

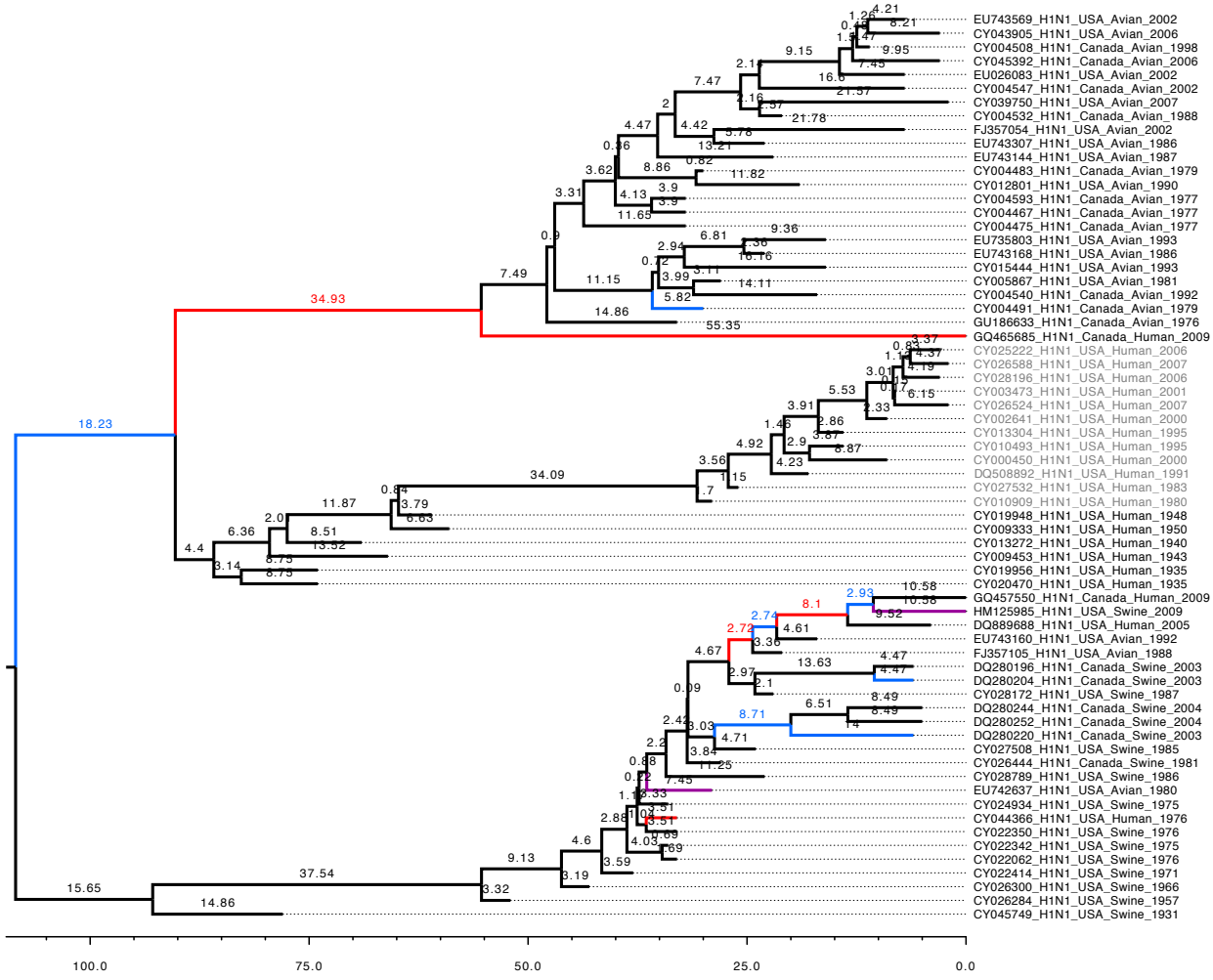

Figure S8: **Timed tree for subtype H1N1 gene M1.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray.

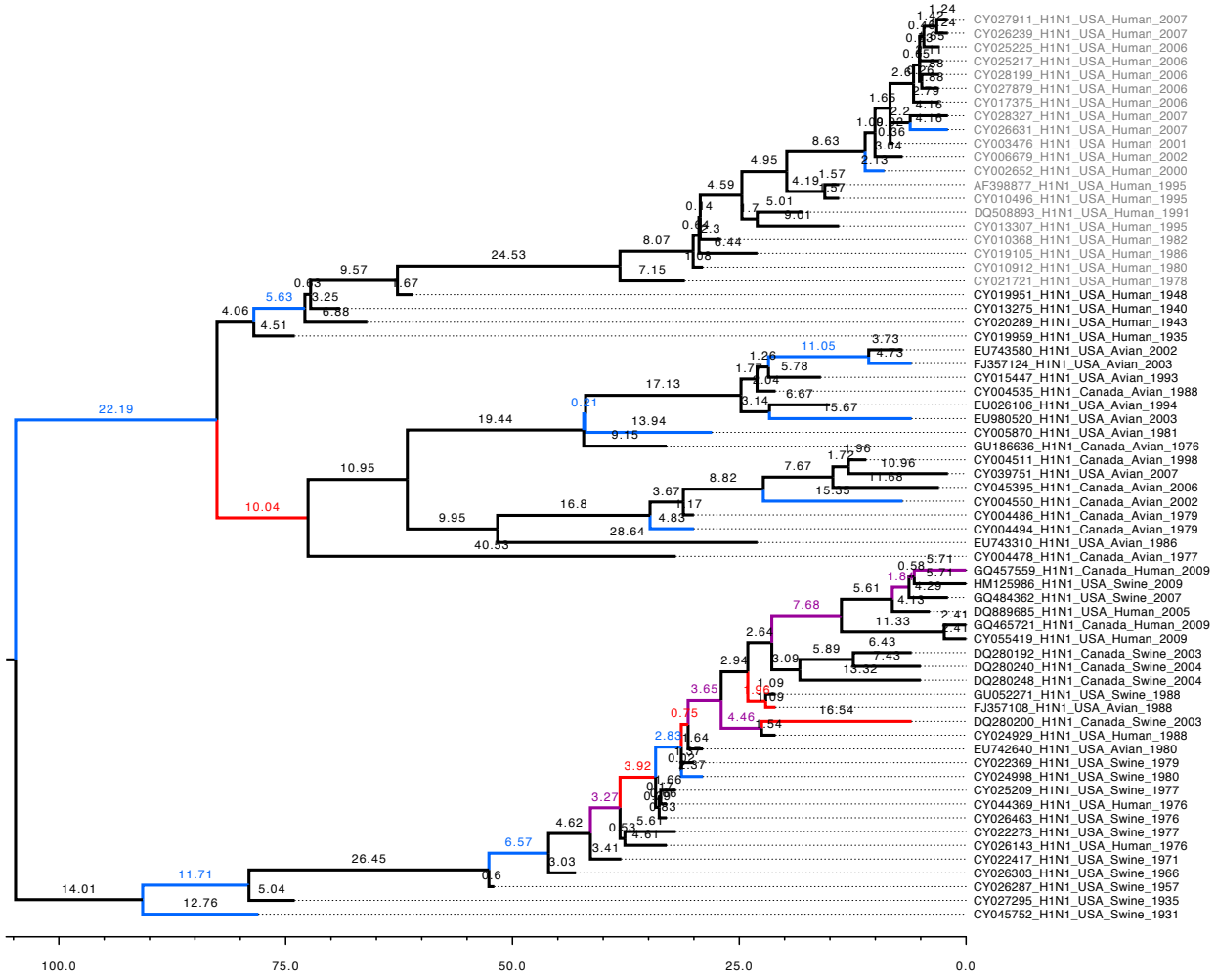

Figure S9: **Timed tree for subtype H1N1 gene NS2.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray.

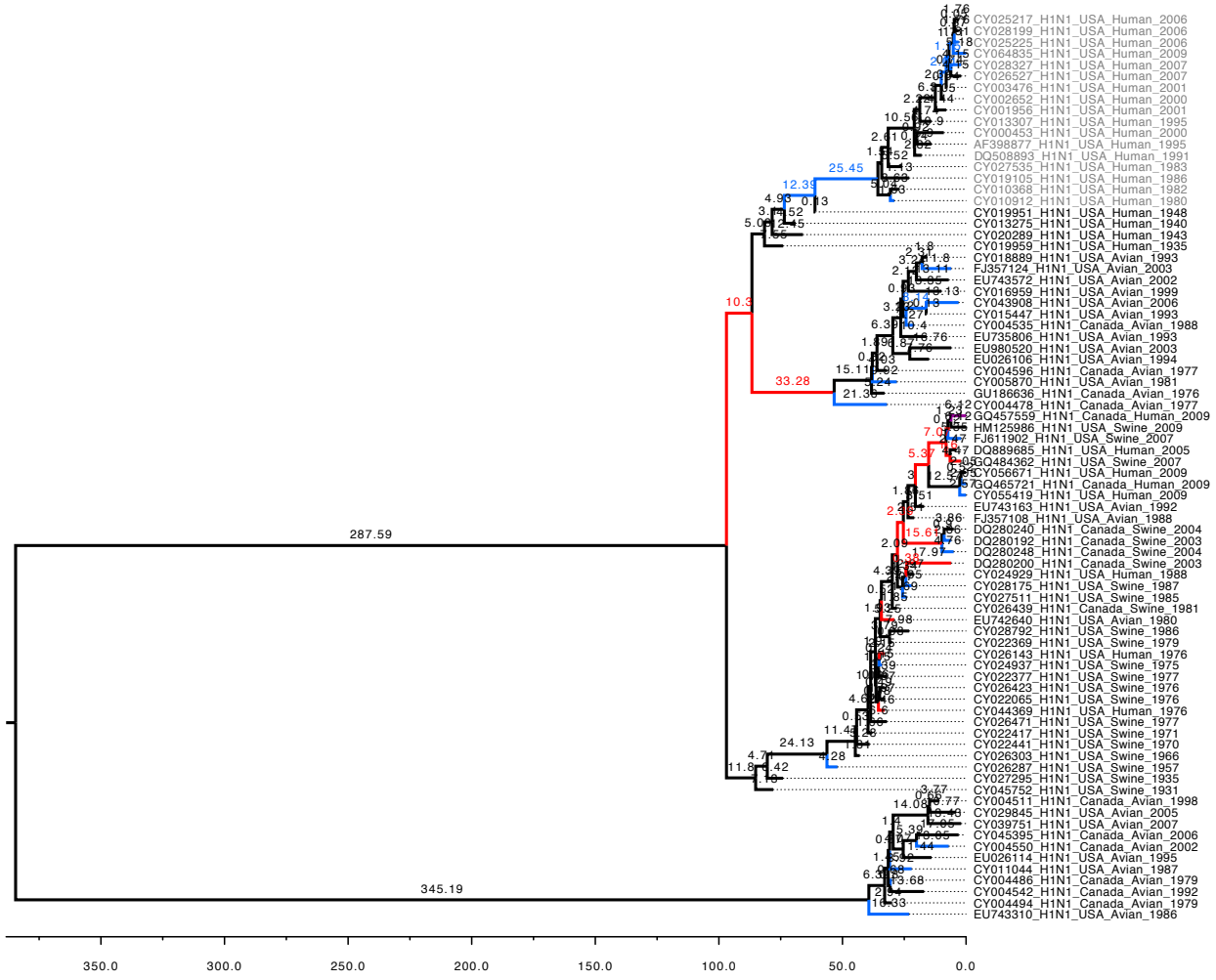

Figure S10: **Timed tree for subtype H1N1 gene NS1.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years. Sequences that reemerged in 1977 after a 20-year absence are indicated in gray.

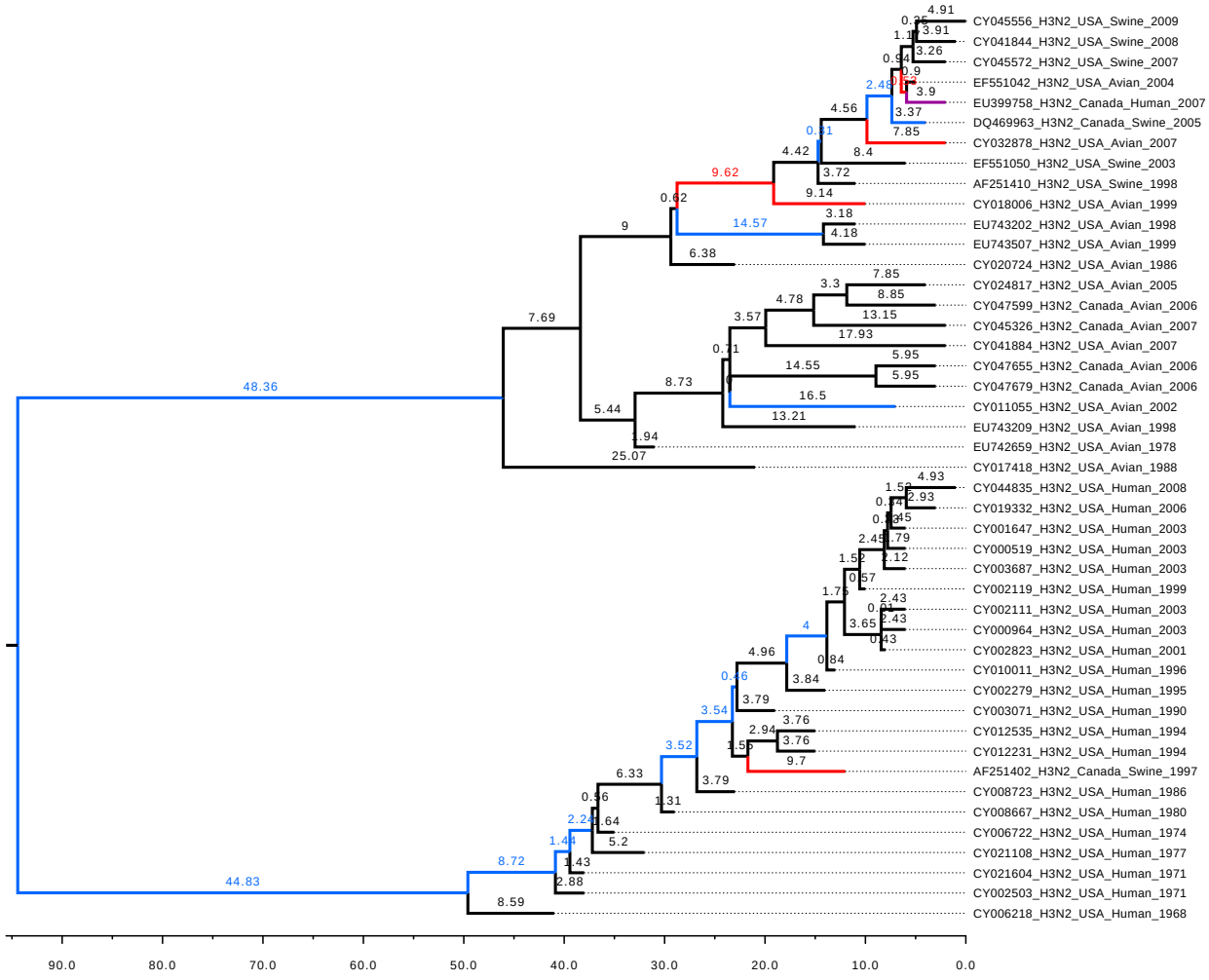

Figure S11: **Timed tree for subtype H3N2 gene PB2.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years.

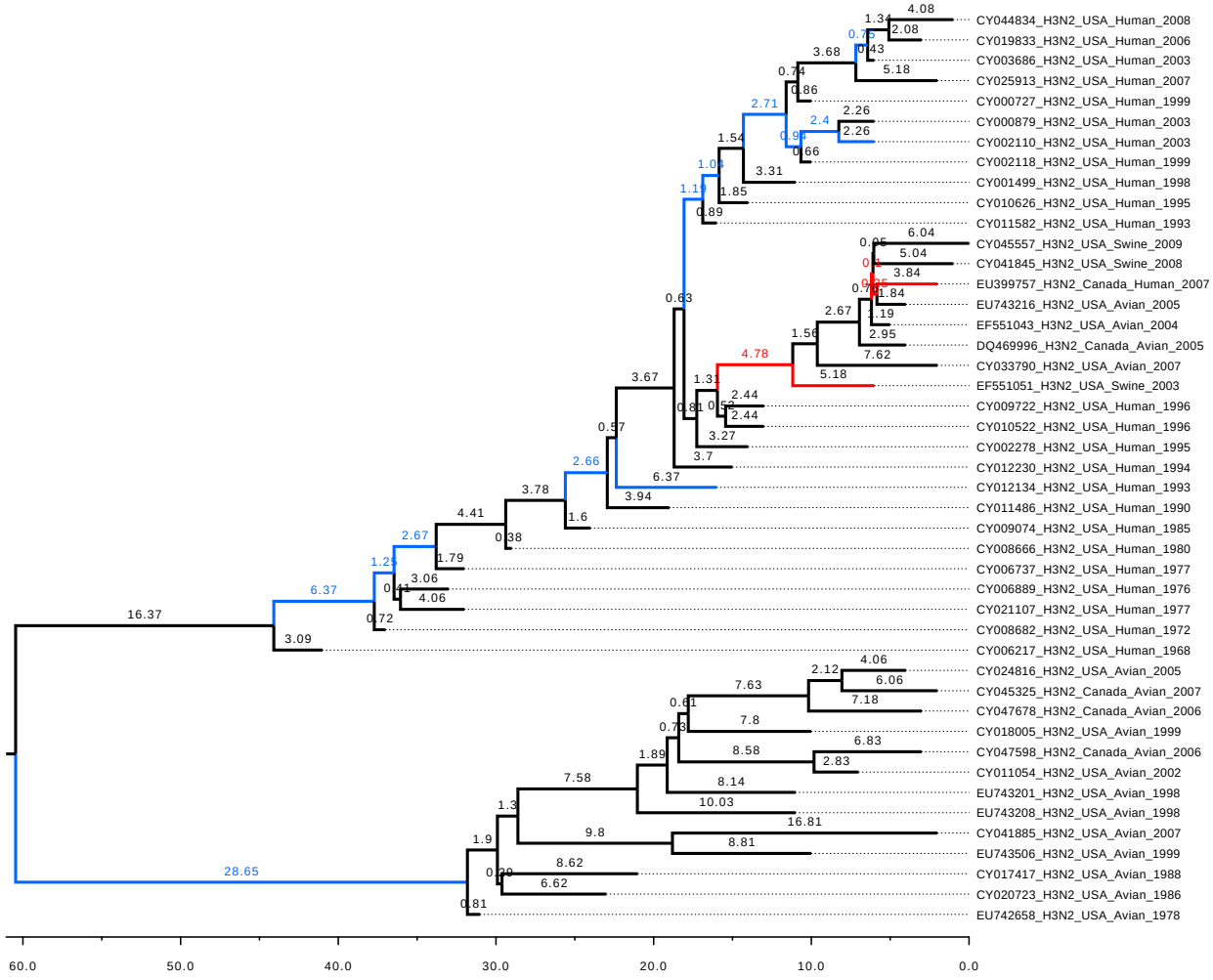

Figure S12: **Timed tree for subtype H3N2 gene PB1.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years.

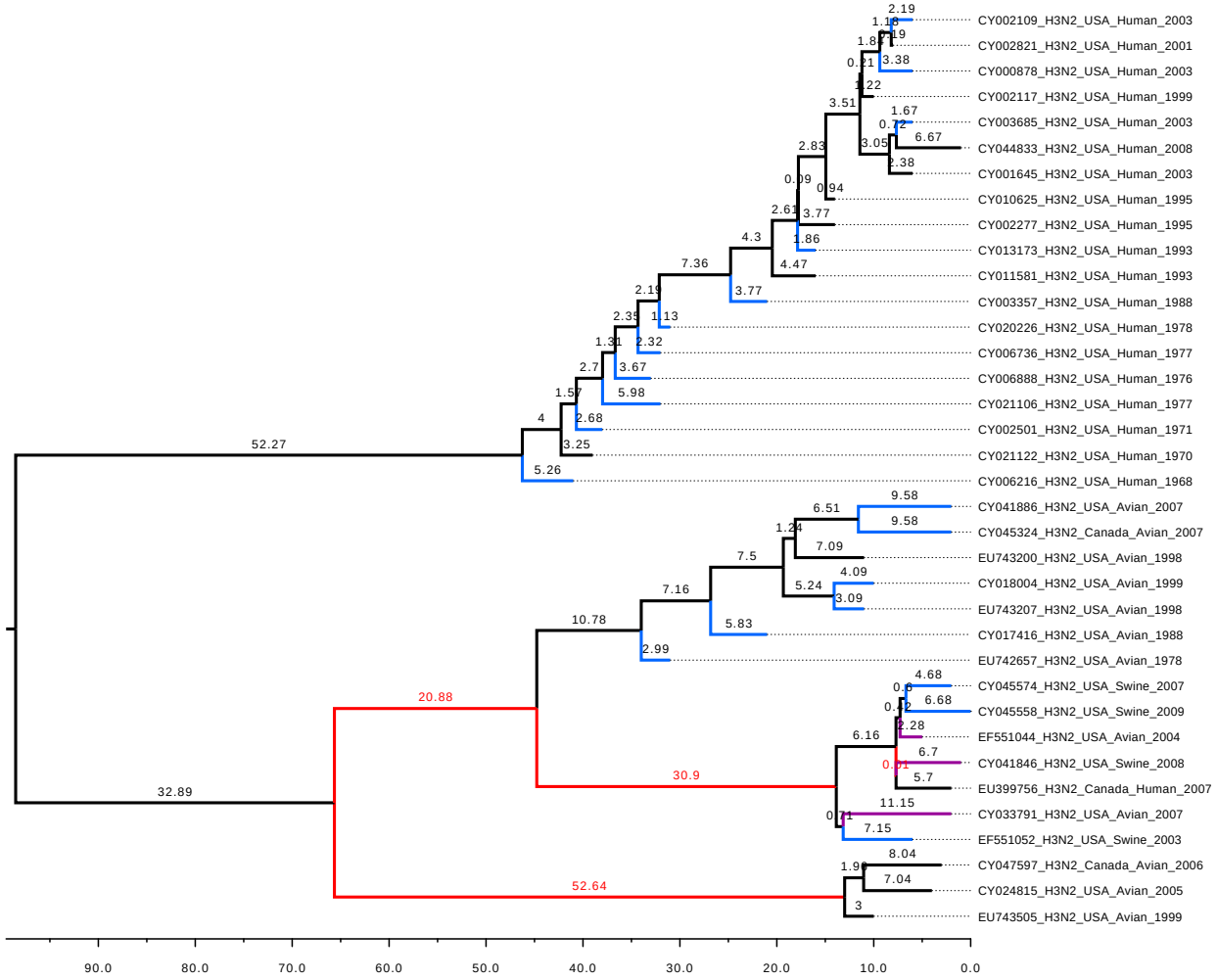

Figure S13: **Timed tree for subtype H3N2 gene PA.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years.

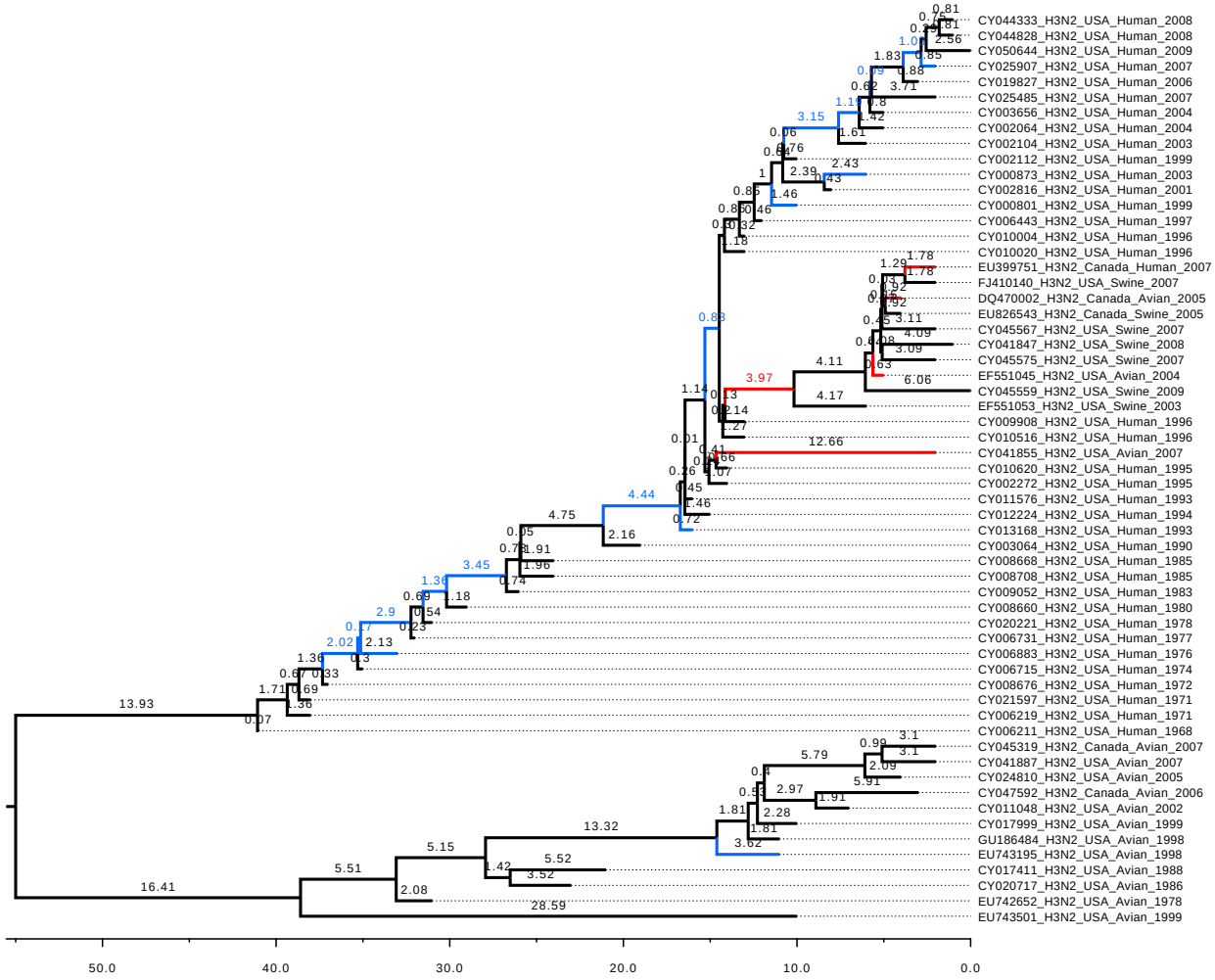

Figure S14: **Timed tree for subtype H3N2 gene HA.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years.

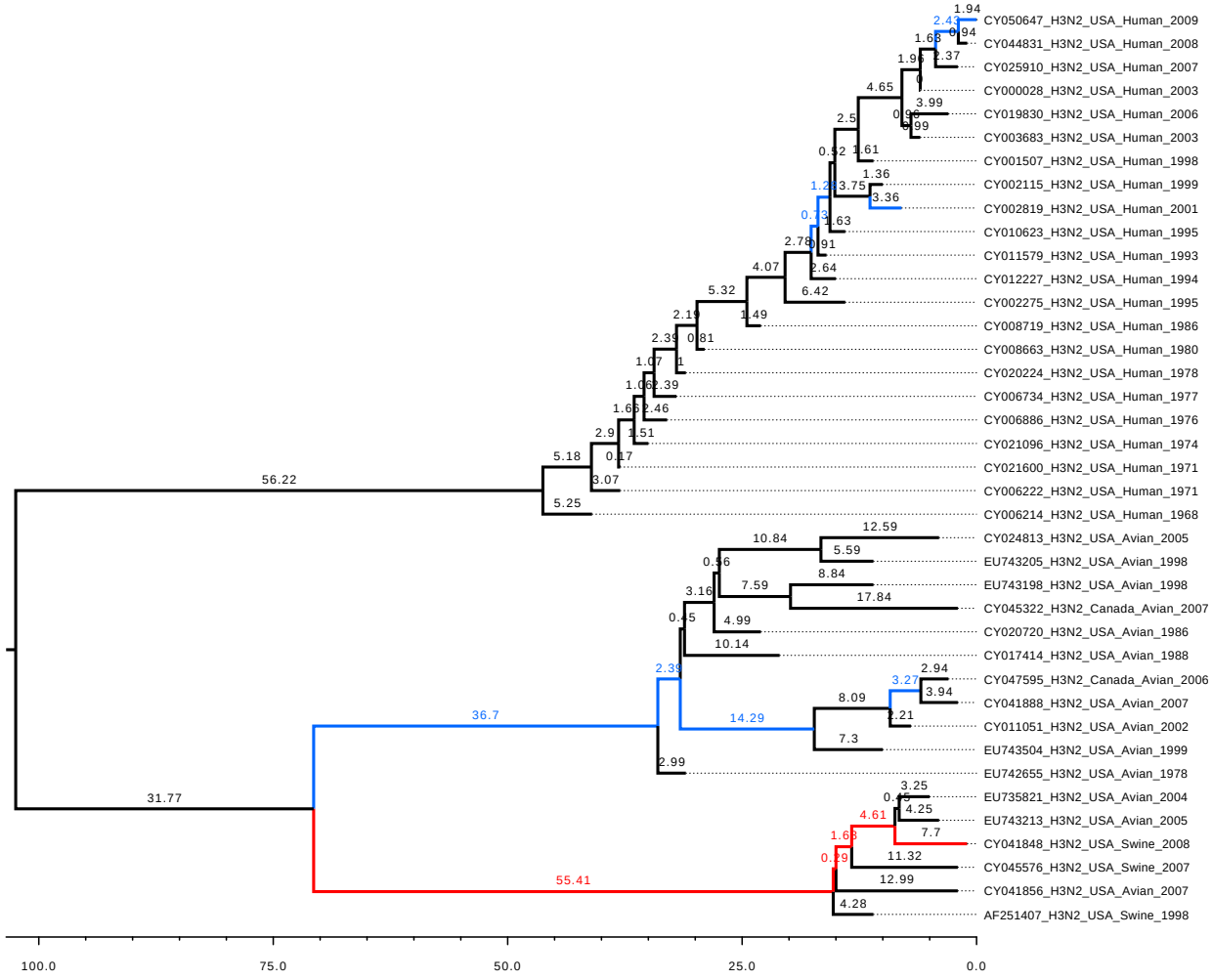

Figure S15: **Timed tree for subtype H3N2 gene NP.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years.

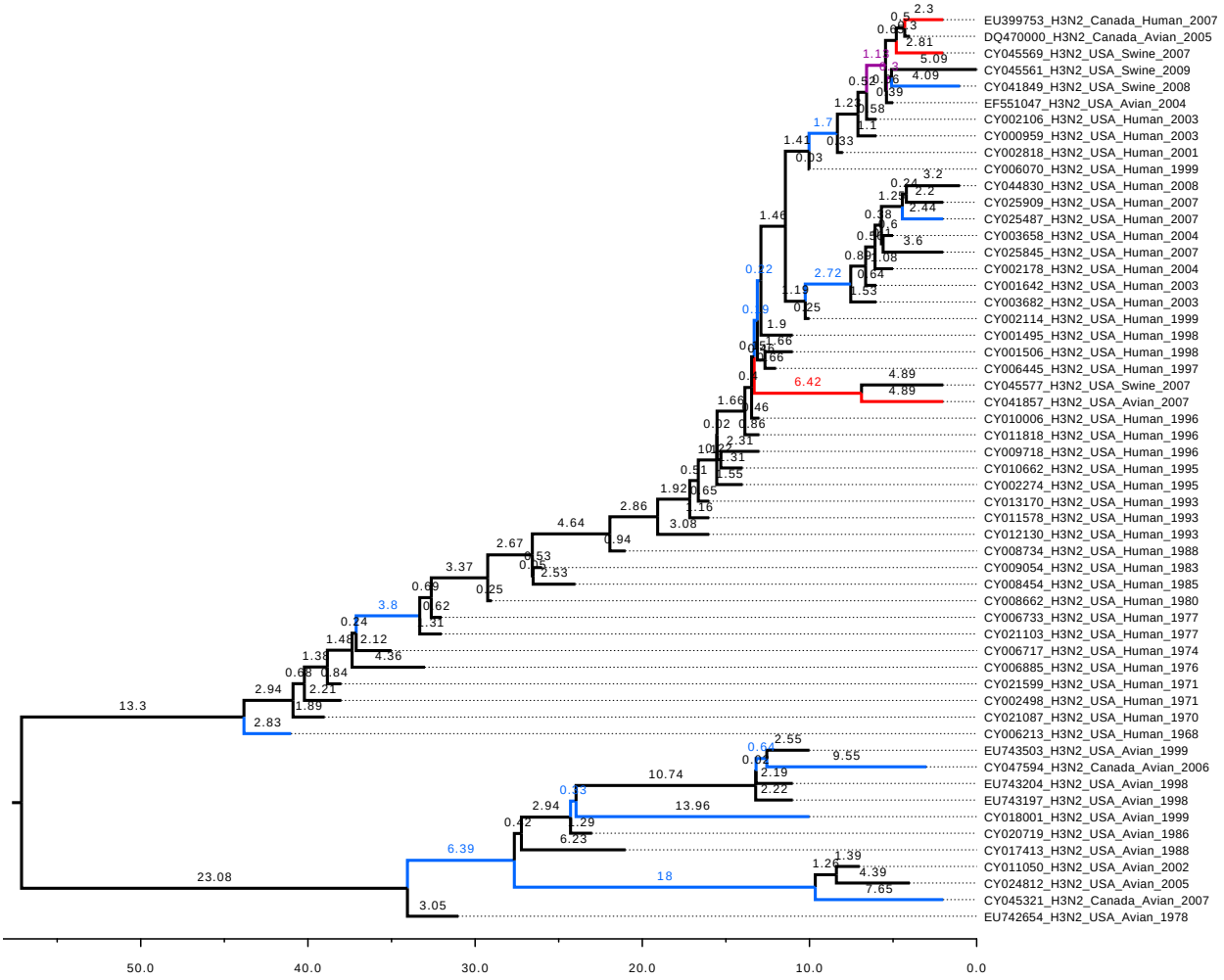

Figure S16: **Timed tree for subtype H3N2 gene NA.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years.

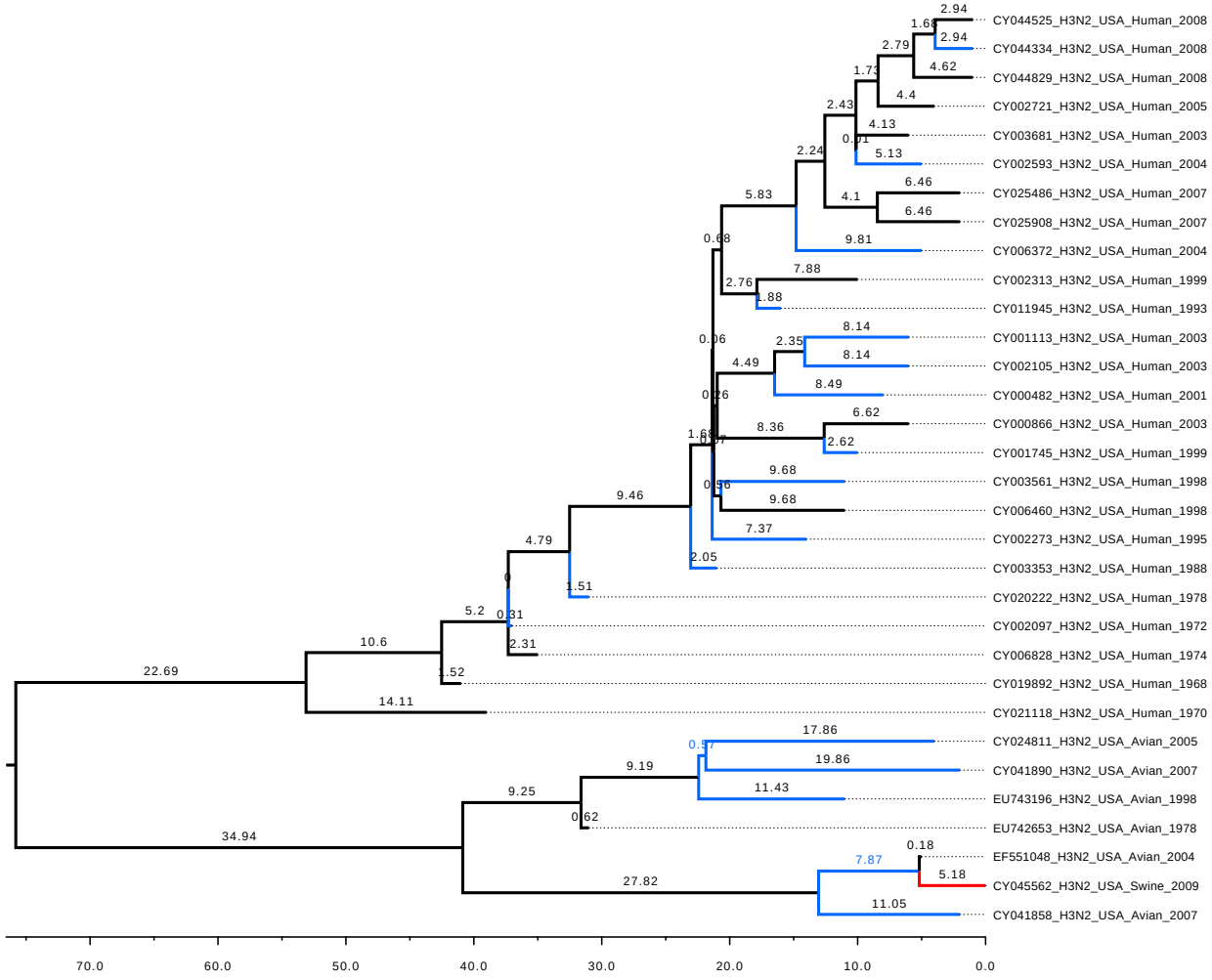

Figure S17: **Timed tree for subtype H3N2 gene M2.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years.

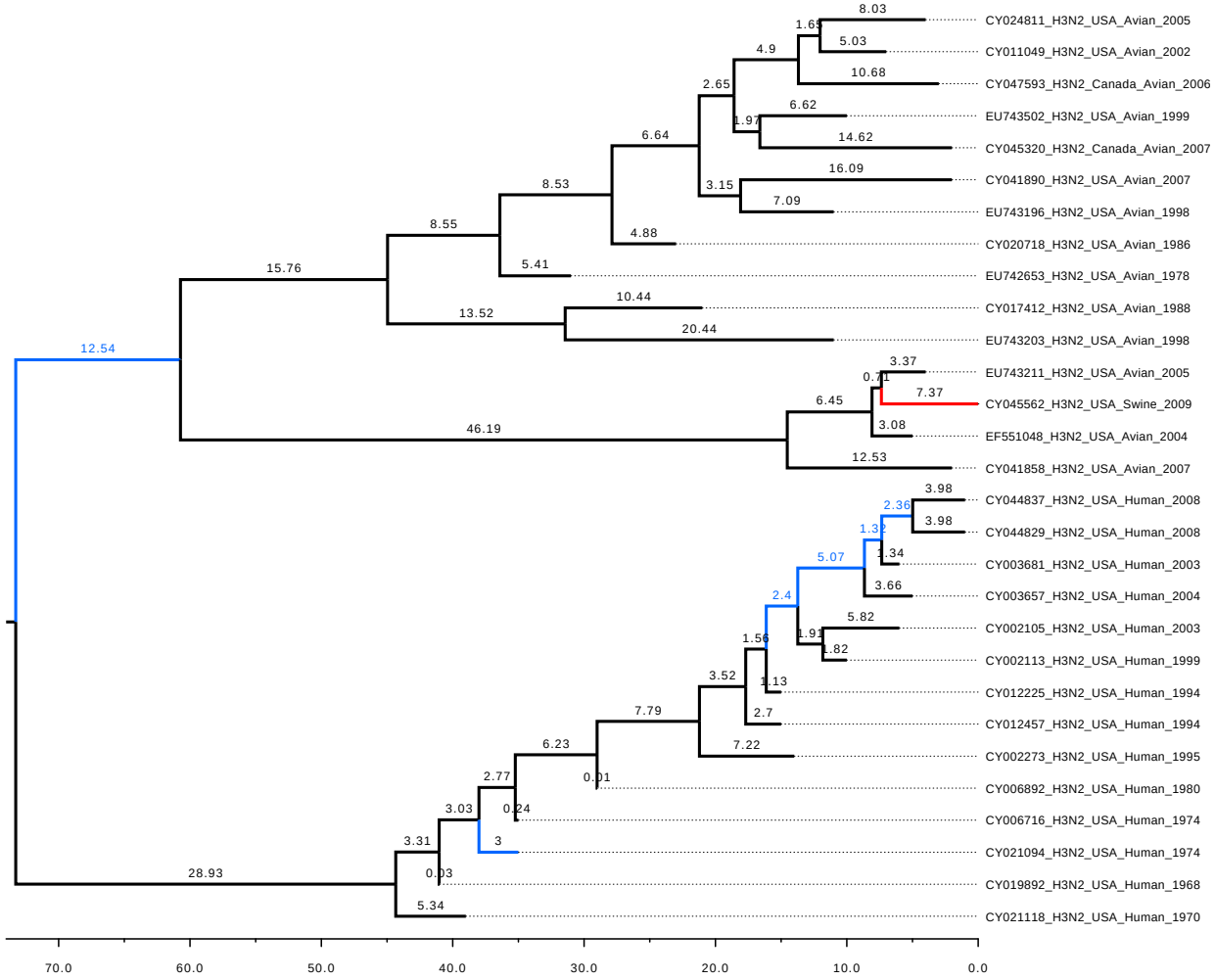

Figure S18: **Timed tree for subtype H3N2 gene M1.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years.

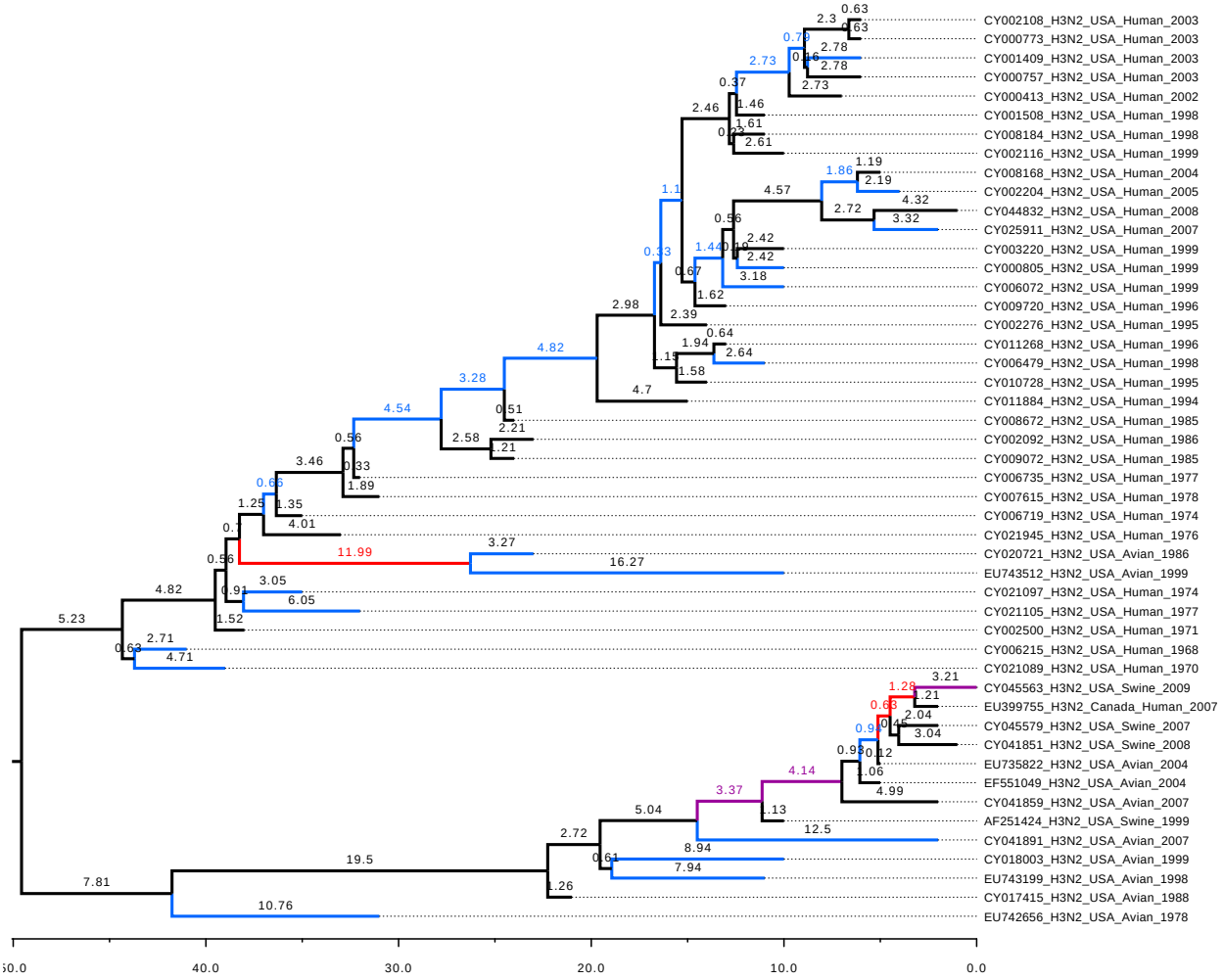

Figure S19: **Timed tree for subtype H3N2 gene NS2.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years.

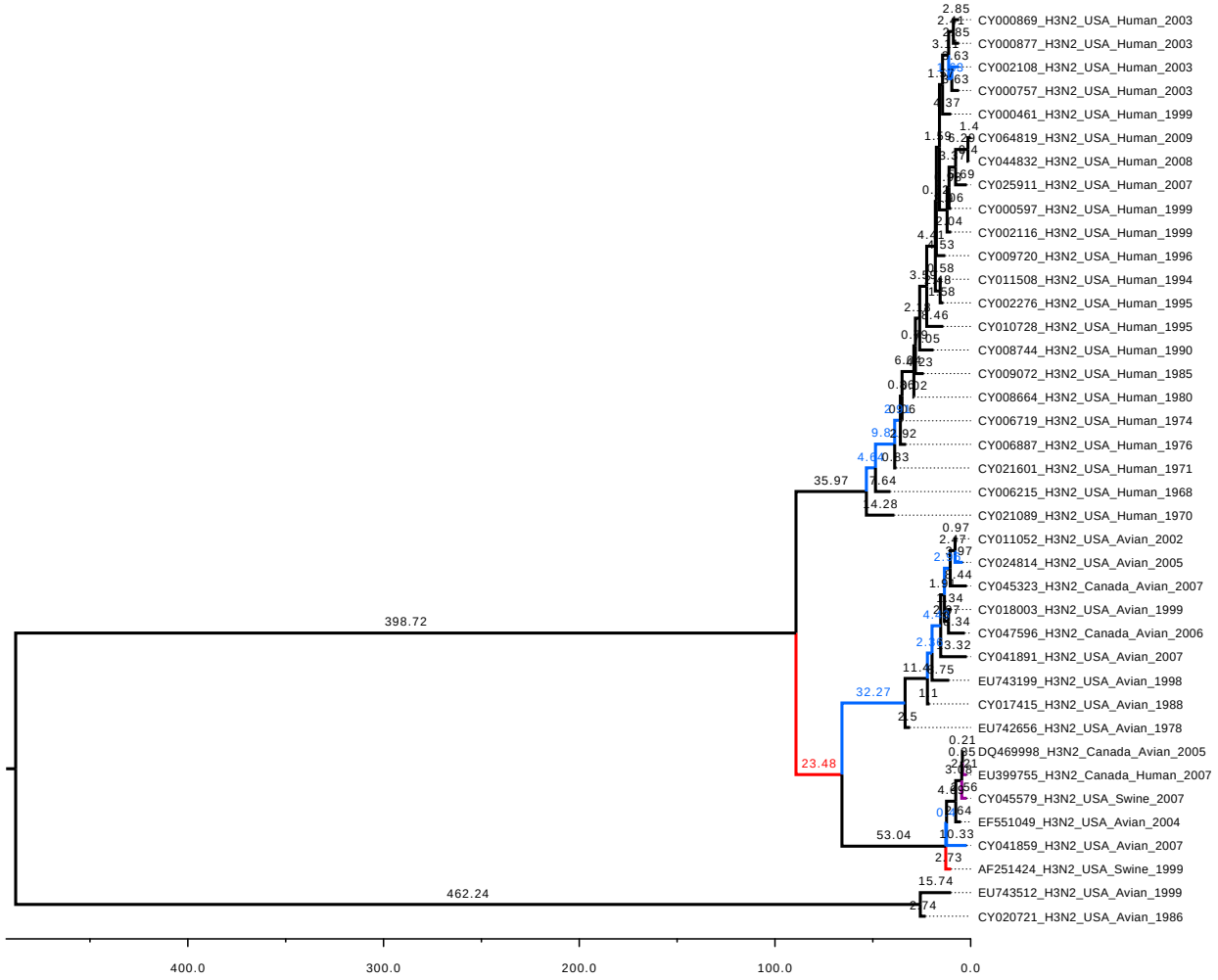

Figure S20: **Timed tree for subtype H3N2 gene NS1.** Branches are color-coded: red for a host-switch event; blue for a change in GC3 cluster; purple for a change in both host and GC3 cluster. Numbers indicate branch lengths in years.

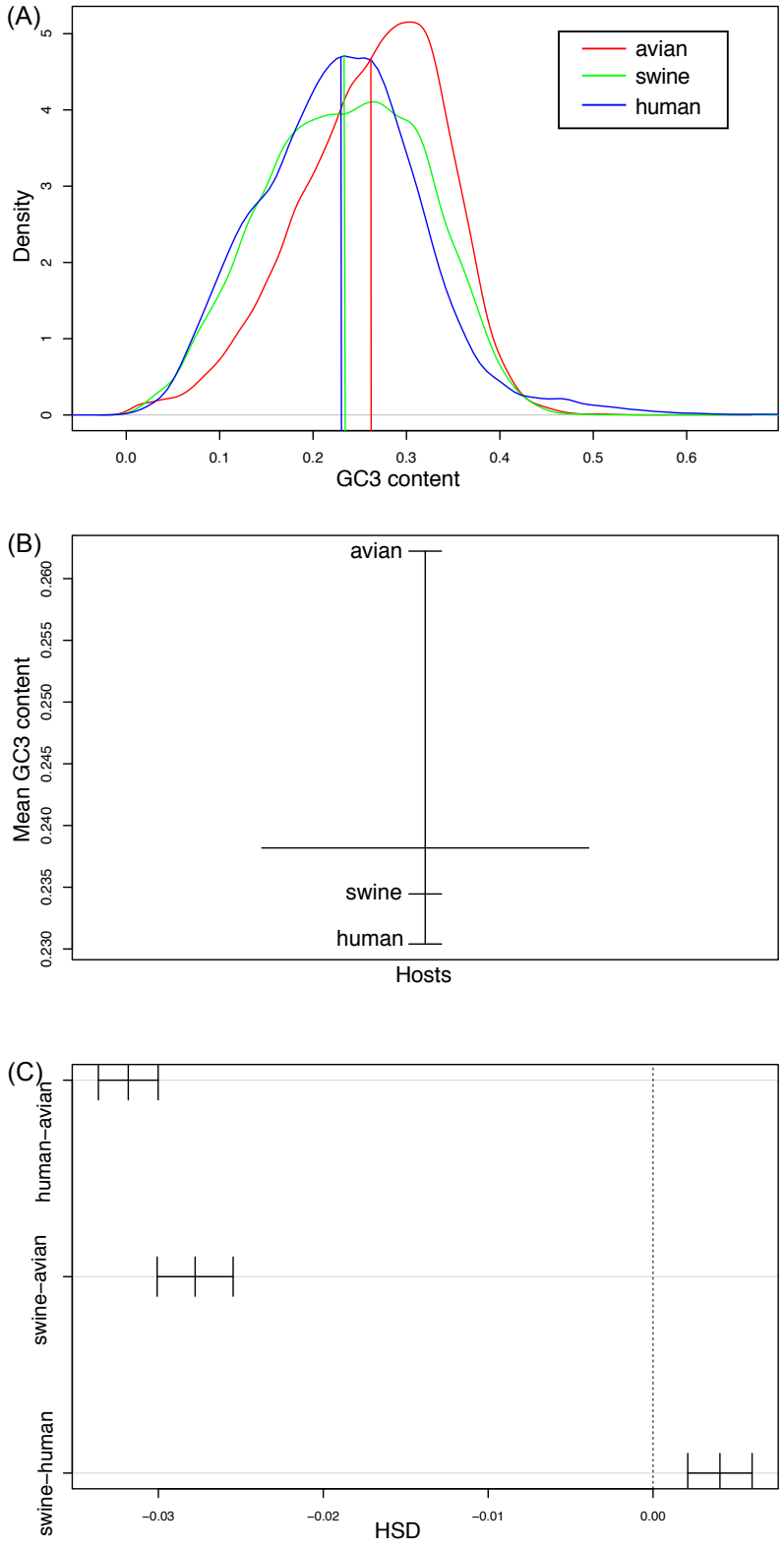

Figure S21: **GC3 composition of the host species.** (A) Density plots of GC3 contents of avian (red), swine (green) and human (blue) hosts; vertical lines represent mean values for each density. (B) Factor effects. (C) Graphical representation of Tukey's Honestly Significant Differences (HSD).

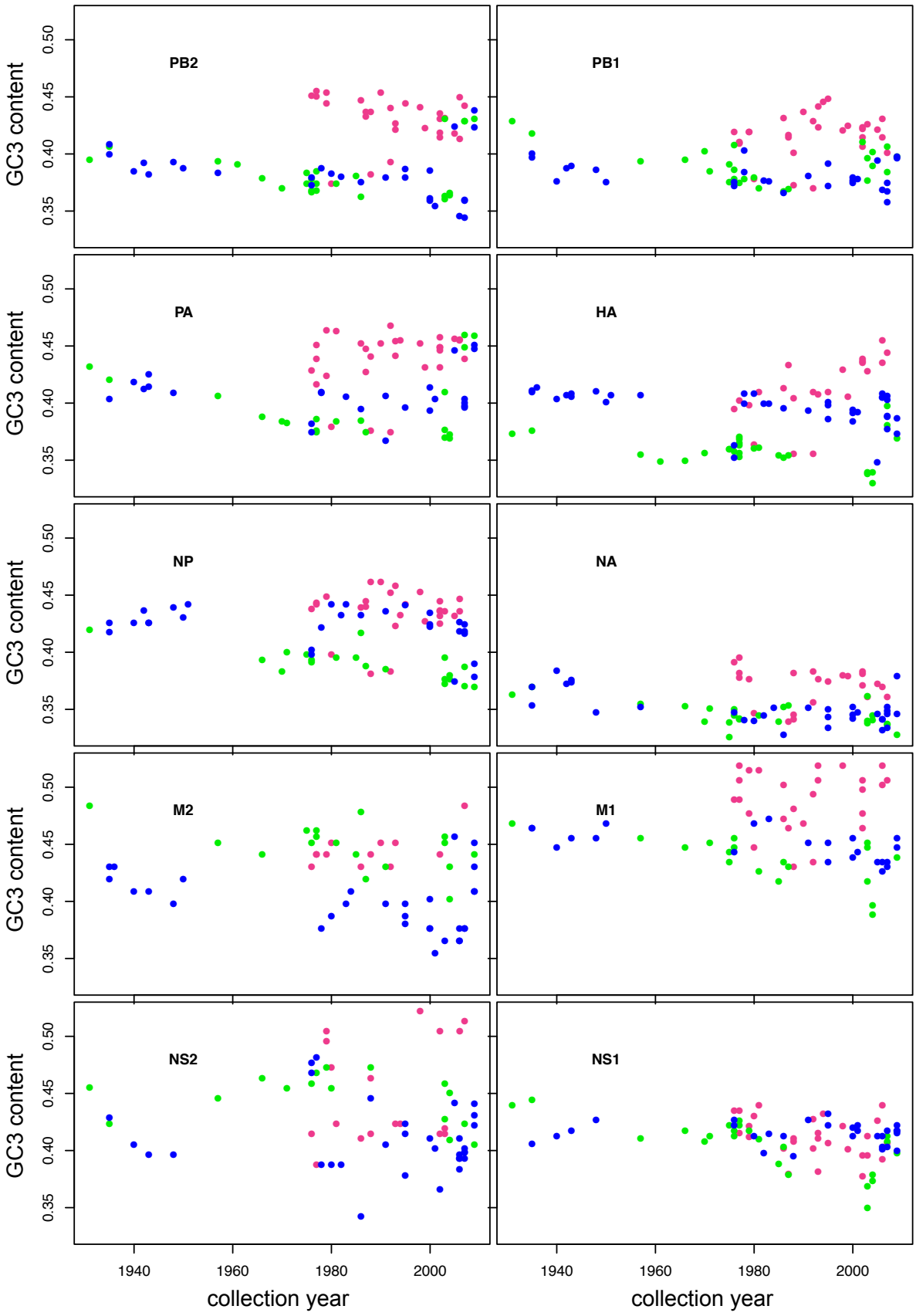

Figure S22: **GC3 composition of H1N1 viruses as a function of collection dates.** Hosts are color-coded: avian in purple, human in blue and swine in green.

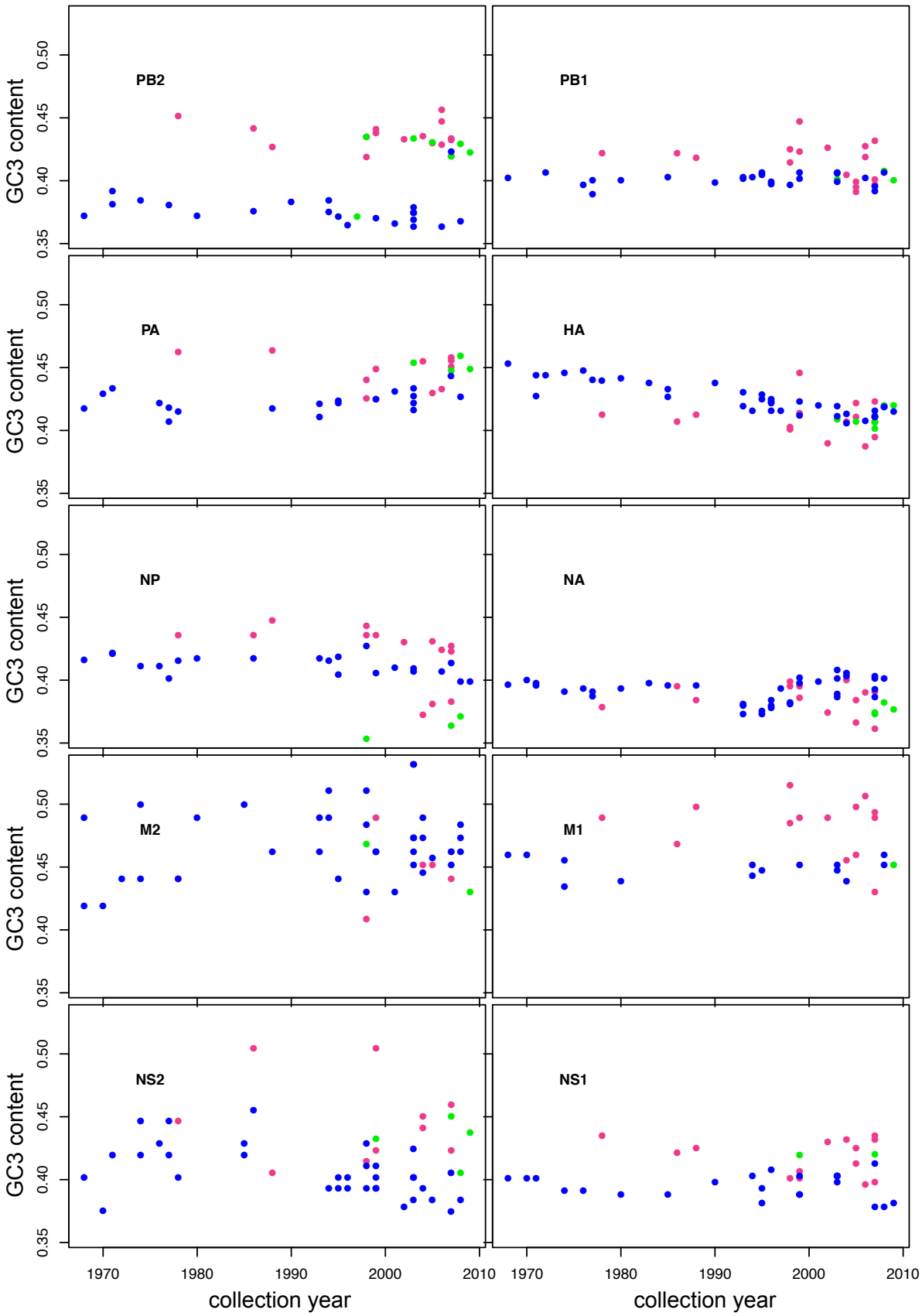

Figure S23: **GC3 composition of H3N2 viruses as a function of collection dates.** Hosts are color-coded: avian in purple, human in blue and swine in green.

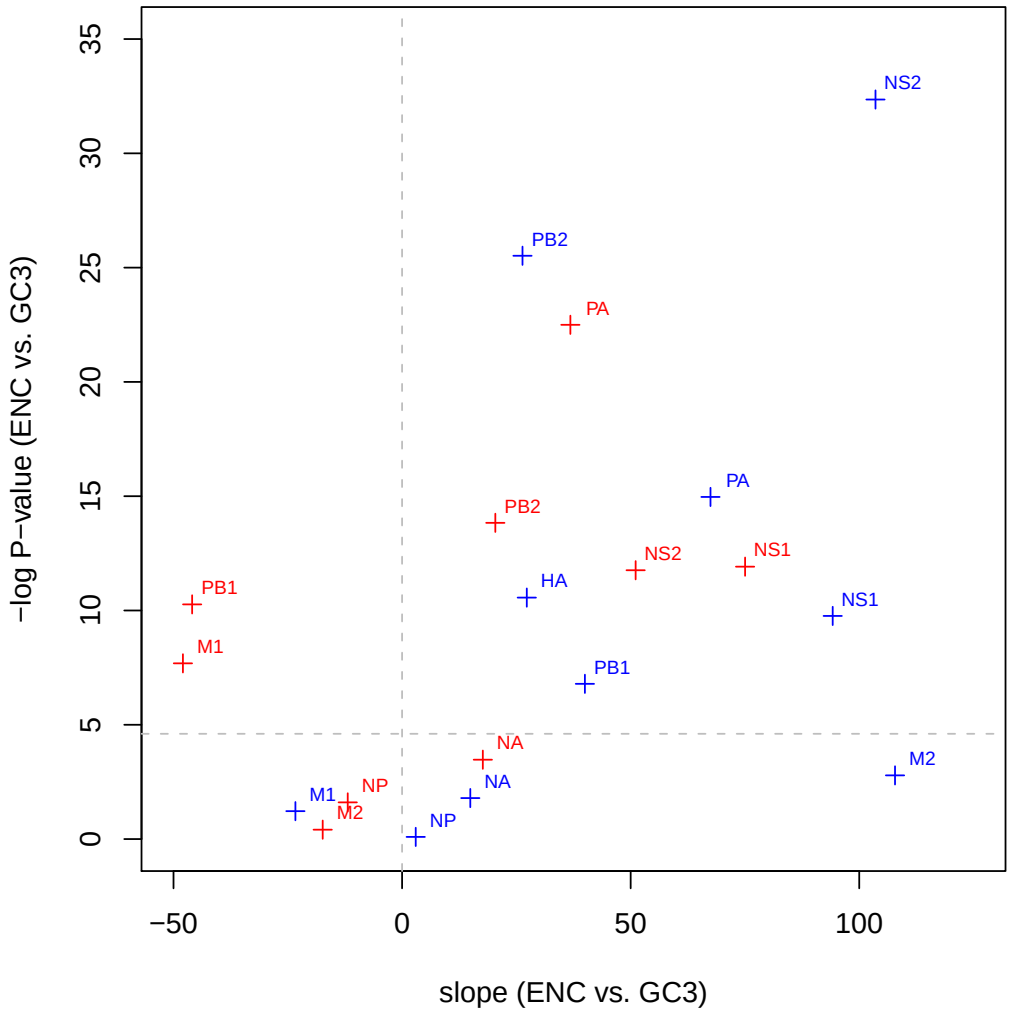

Figure S24: **Significance of the linear fit between effective number of codons and GC3 content.** H1N1 viruses are in red; H3N2 viruses are in blue. The significance level is taken at 1% (dashed horizontal line in gray). HA in H1N1 has a P-value of 0, and therefore does not appear on the graph.

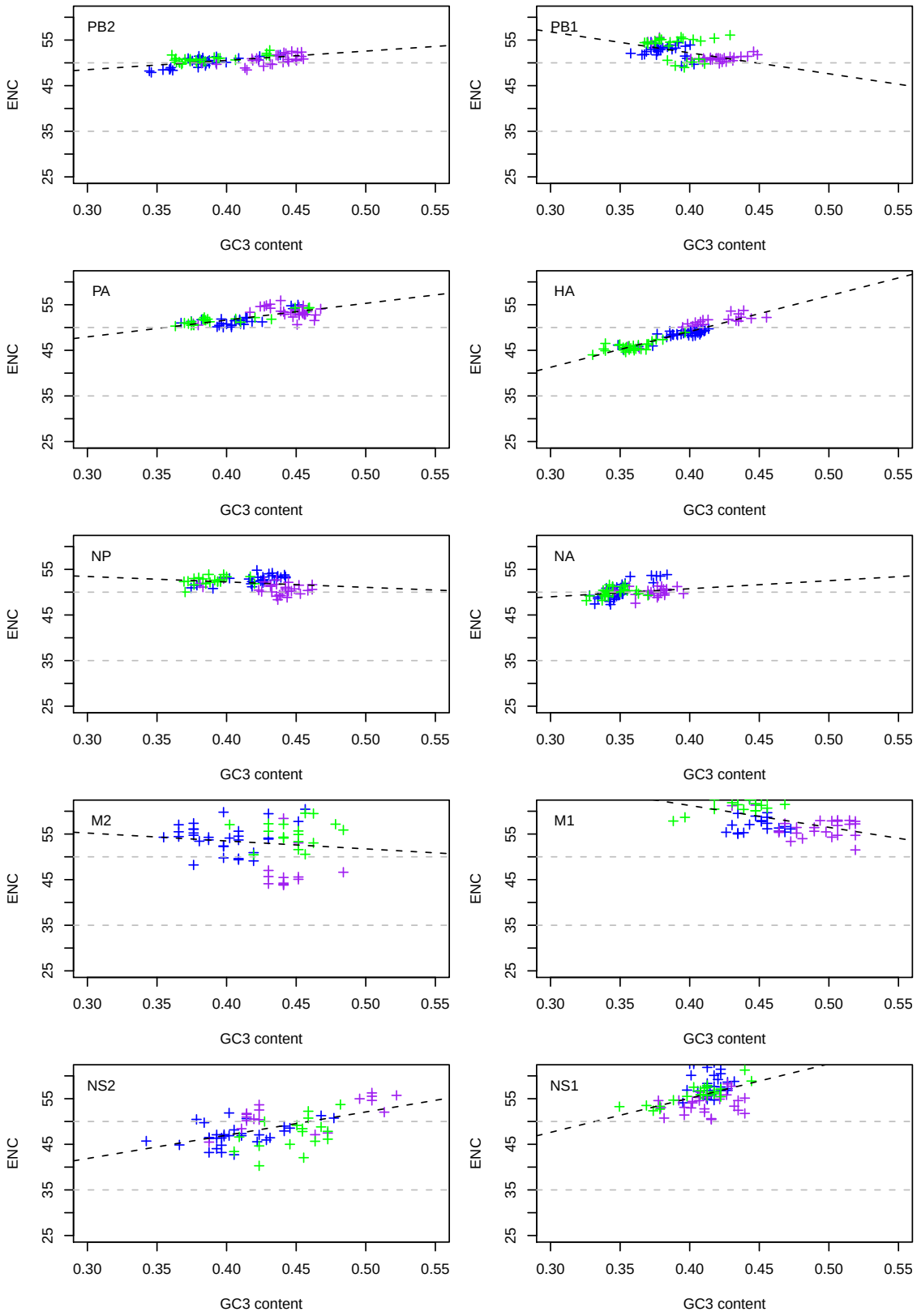

Figure S25: **Effective number of codons (ENC) as a function GC3 content for H1N1 viruses.** Hosts are color-coded: avian in purple, human in blue and swine in green. Gray horizontal lines represent ENC cutoffs at 35 and 50. The linear fit is represented as a black dashed line.

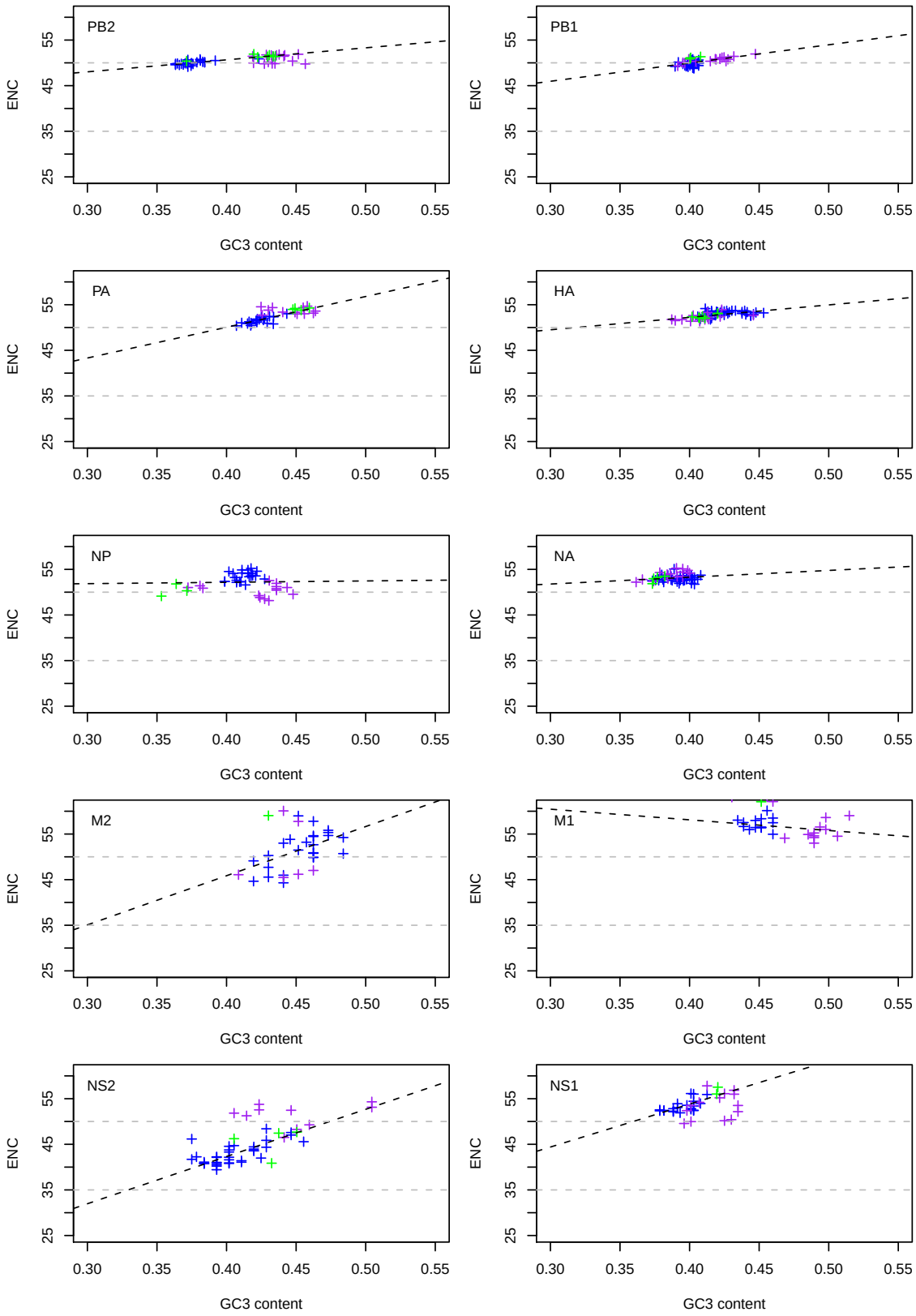

Figure S26: **Effective number of codons (ENC) as a function GC3 content for H3N2 viruses.** Hosts are color-coded: avian in purple, human in blue and swine in green. Gray horizontal lines represent ENC cutoffs at 35 and 50. The linear fit is represented as a black dashed line.

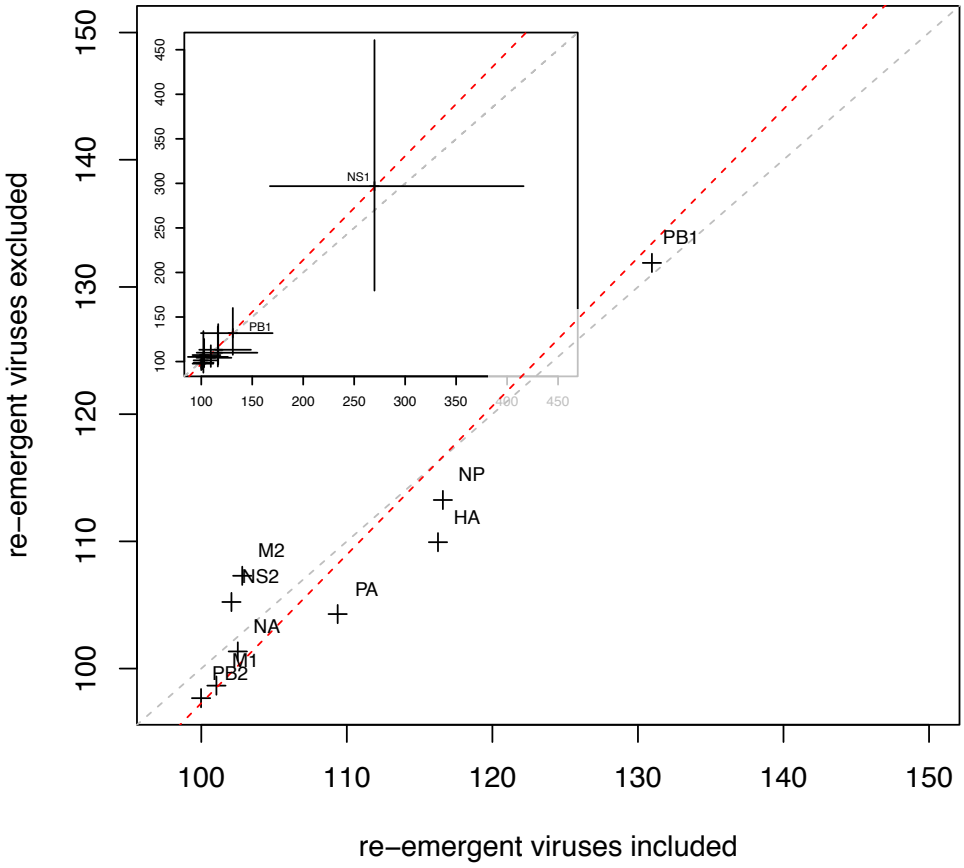

Figure S27: **Comparison of root age for H1N1 viruses with and without re-emergent strains.** The gray line represents the first bisector (line of equation $y = x$), while the red line represents the linear fit to the data. Scale on both axes is in years before 2009. Insert shows the results for all ten 'canonical' genes, with bars representing the 95% Highest Posterior Densities.

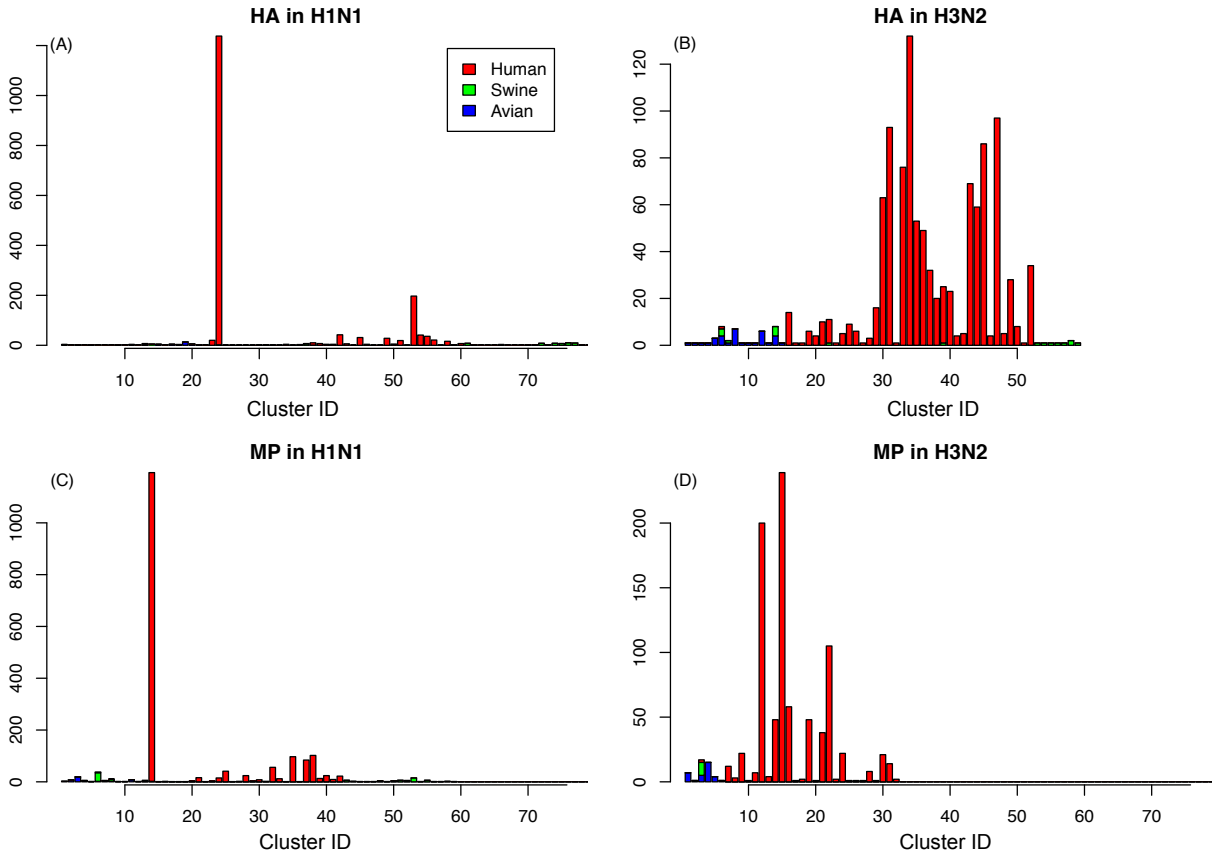

Figure S28: **Cluster composition for select H1N1 and H3N2 genes.** Results of sequences clustering are shown for the fastest evolving gene, HA, for (A) H1N1 and (B) H3N2 viruses as well as for the slowest evolving gene, M1, for (C) H1N1 and (D) H3N2. Hosts are represented in red (human), green (swine) and blue (avian). Cluster IDs are arbitrary; their largest number represent the final number of sequences in the dating analyses.