Published in final edited form as:

Infect Genet Evol. 2011 December ; 11(8): 2125–2132. doi:10.1016/j.meegid.2011.07.002.

Evolutionary dynamics of influenza A nucleoprotein (NP) lineages revealed by large-scale sequence analyses

Jianpeng Xu¹, Mary C. Christman¹, Ruben O. Donis², and Guoqing Lu^{1,*}

¹ Department of Biology, University of Nebraska at Omaha, Omaha, NE 68182, the United States
jianpenxu@unomaha.edu for JX; mchristman@unomaha.edu for MC

² Influenza Division, Molecular Virology and Vaccines Branch, Centers for Disease Control and Prevention, Atlanta, Georgia, USA rvd6@cdc.gov

Abstract

Influenza A viral nucleoprotein (NP) plays a critical role in virus replication and host adaptation, however, the underlying molecular evolutionary dynamics of NP lineages are less well-understood. In this study, large-scale analyses of 7,517 NP nucleotide sequences revealed eight distinct evolutionary lineages, including three host-specific lineages (human, classic swine and equine), two cross-host lineages (Eurasian avian-like swine and swine-origin human pandemic H1N1 2009) and three geographically isolated avian lineages (Eurasian, North American and Oceanian). The average nucleotide substitution rate of the NP lineages was estimated to be 2.4×10^{-3} substitutions per site per year, with the highest value observed in pandemic H1N1 2009 (3.4×10^{-3}) and the lowest in equine (0.9×10^{-3}). The estimated time of most recent common ancestor (TMRCA) for each lineage demonstrated that the earliest human lineage was derived around 1906, and the latest pandemic H1N1 2009 lineage dated back to Dec 17, 2008. A marked time gap was found between the times when the viruses emerged and were firstly sampled, suggesting the crucial role for long-term surveillance of newly emerging viruses. The selection analyses showed that human lineage had six positive selection sites, whereas pandemic H1N1 2009, classical swine, Eurasian avian and Eurasian swine had only one or two sites. Protein structure analyses revealed several positive selection sites located in epitope regions or host adaptation regions, indicating strong adaptation to host immune system pressures in influenza viruses. Along with previous studies, this study provides new insights into the evolutionary dynamics of influenza A NP lineages. Further lineage analyses of other gene segments will allow better understanding of influenza A virus evolution and assist in the improvement of global influenza surveillance.

Keywords

influenza; nucleoprotein (NP); lineage; substitution rate; TMRCA; selection

© 2011 Elsevier B.V. All rights reserved.

* Correspondence: Department of Biology, University of Nebraska at Omaha Omaha, NE 68182-0040 Tel: 1-402-5543195 Fax: 1-402-5543532 glu3@unomaha.edu .

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Influenza A virus, which can infect a wide range of avian and mammalian species, has been considered an extremely important pathogen of respiratory infections (Webster et al., 1992). The virus can cause moderate to severe epidemics annually and catastrophic pandemics sporadically (Guan et al., 2010). The occurrence of pandemic H1N1 2009 attracted great public attention concerning the global burden of morbidity and mortality caused by influenza viruses (Garten et al., 2009). The influenza A virus is a negative strand RNA virus with eight gene segments, each of which encodes either one or two proteins. The subtype of influenza A virus is determined by the antigenicity of two surface glycoproteins, namely, hemagglutinin (HA) and neuraminidase (NA). Currently, 16 HA subtypes (H1-H16) and 9 NA subtypes (N1-N9) have been classified within influenza A viruses (Fouchier et al., 2005). The subtypes currently circulating in the human population are H1N1 and H3N2. More than 100 of the possible 144 influenza A virus HA - NA combinations are maintained in wild bird populations, with many variants found in each subtype (Munster et al., 2007).

Influenza A virus exhibits rapid evolution and complicated molecular dynamics due to diverse hosts, high mutation rates, and rapid replication (Holmes, 2010). The virus has also shown the propensity to escape immunity because of continuous antigenic drift, during which the viral protein epitopes are altered by non-synonymous mutations, as a result of interactions with the host immune system (Bush et al., 1999; Ferguson et al., 2003; Smith et al., 2004). In addition, the influenza virus can cross species barriers due to rapid mutation, genetic drift, genome reassortment, or simply transmitted from one host to another (Holmes, 2010; Kuiken et al., 2006; Pensaert et al., 1981). Each of the influenza viral genes plays a significant role in viral interactions with hosts and subsequent infection; therefore, understanding the evolution of each viral gene can provide new insights into the epidemiological dynamics of influenza viruses (Fourment et al., 2010; Furuse et al., 2009). This will be useful in predicting the genetic basis and periodicity of future influenza epidemics and pandemics (Pybus and Rambaut, 2009).

The nucleoprotein (NP) gene encodes a protein with around 500 amino acids; it plays an important role in assembly and budding of influenza virus and has a putative role in host range (Ruigrok et al., 2010; Snyder et al., 1987). The primary function of NP is to bind with the viral RNA segments to form the nucleocapsid of a virus particle. RNA hybridization techniques were used to investigate the genetic diversity and grouping of NP sequences, which revealed five RNA hybridization groups (Bean, 1984). Sequence analyses of the NP gene using the maximum-parsimony algorithm supports the hypothesis that NP is involved in the maintenance of a host-specific gene pool (Gorman et al., 1990). In addition, phylogenetic analysis of 89 NP nucleotide sequences demonstrated that H1N1 human and classical swine virus share a common ancestor (Gorman et al., 1991). The phylogenetic analyses of 436 NP sequences provides a view of phylogenetic diversity and distribution of lineages; however, some lineages were assigned based upon the information of as few as two sequences (Chen et al., 2009). On the other hand, the phylodynamics parameters of the NP gene (i.e., evolutionary rates and TMRCA) were analyzed in genomic studies but limited to certain specific hosts (Chen and Holmes, 2006, 2010). Our large-scale sequence analyses of the influenza A viral NP gene were conducted to fill the above gaps.

2. Materials and Methods

2.1 Sequence data

A complete set of nucleoprotein (NP) nucleotide sequences, together with the information of host, subtype, isolation time, and isolation place, were downloaded from GenBank (Bao et al., 2008). After excluding the sequences from laboratory strains or duplicate strains, a total

of 7,517 sequences longer than 1,440 nucleotides were obtained. The identical sequences were removed using a Perl script, which resulted in 5,094 sequences for further analysis.

2.2 Phylogenetic analysis

Sequences were aligned with MUSCLE (Edgar, 2004). The alignments were adjusted manually using TranslatorX (Abascal et al., 2010) based upon corresponding protein sequences. Sequences were tested for recombination using RDP3 (Martin et al., 2010); no definitive recombinant was detected. Phylogenetic analysis was conducted using the maximum-likelihood (ML) method in RAxML (Stamatakis et al., 2005). RAxML uses rapid algorithms for bootstrap and maximum likelihood searches and is considered one of the fastest and most accurate phylogeny programs for large datasets. Analyses of 1000 bootstrap replicates were performed using GTRGAMMA, i.e., GTR model of nucleotide substitution with Gamma model of rate heterogeneity. All the analyses were conducted on a Supercomputer cluster at the University of Nebraska Holland Computing Center (<http://hcc.unl.edu>). A set of Perl scripts were written by us to facilitate their computational analyses.

2.3 Classification and nomenclature of lineages

Lineages and major clusters were determined based upon the groupings with strong bootstrap support in the ML tree. Additional information such as hosts, geographic regions and circulation years were also considered in the classification. The trees were visualized and color-coded with FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>). We used the same lineage nomenclature as we described previously: Roman alphabetic letters such as A, B and C to represent different lineages (Lu et al., 2007).

2.4 Measurement of selection pressures

To determine selection pressure acting on NP lineages, we estimated the ratio of non-synonymous (d_N) to synonymous (d_S) substitutions per site (ratio d_N/d_S) for each lineage, using all the sequences included in this study. Positively selected codons were detected using the single likelihood ancestor counting (SLAC), fixed effects likelihood (FEL) and internal fixed effects likelihood (IFEL) methods with a significance level of 0.05. In the SLAC method, the nucleotide and codon model parameter estimates are used to reconstruct the ancestral codon sequences at the internal nodes of the tree. The single most likely ancestral sequences are then fixed as known variables, and applied to infer the expected number of non-synonymous or synonymous substitutions that have occurred along each branch, for each codon position. The FEL method is based on maximum-likelihood estimates. The FEL method estimates the ratio of non-synonymous to synonymous substitutions on a site-by-site basis for the entire tree or only the interior branches (IFEL). These methods are implemented in the HYPHY package (Pond et al., 2005). Positive selection sites were mapped on the NP protein structure using Molecular Operating Environment (MOE) (Ye et al., 2006).

2.5 Substitution rate and time of most recent common ancestor (TMRCA) analyses

The substitution rate and the time of most recent common ancestor (TMRCA) of each lineage were estimated using a Bayesian Markov Chain Monte Carlo (MCMC) method as implemented in the BEAST version 1.5.4 package (Drummond and Rambaut, 2007). For computational tractability, samples from the complete datasets of size 120 - 200 sequences were carefully selected by removing sequences collected from the same location and at the same point in time (Felsenstein, 2006). In all cases, the GTR + Γ_4 nucleotide substitution model was employed as this was the best-fit model supported by Modeltest (Posada and Crandall, 1998). For each analysis the Bayesian skyline coalescent model was used, as it

describes the fluctuating population dynamics characteristics of influenza virus (Rambaut et al., 2008).

Three clock models were compared statistically for each dataset using a Bayes factor test in the Tracer program (Suchard et al., 2001): a strict clock, an uncorrelated lognormal relaxed clock (UCLD) and an uncorrelated exponential relaxed clock (UCED) (Drummond et al., 2006). The UCED model was found to provide the best fit for all lineages (Bayes factor > 30). We thus used this model for the estimation of evolutionary dynamic parameters. In each case, MCMC chains were run for sufficient time to achieve convergence, with uncertainty in parameter estimates reflected in the 95% highest probability density (HPD). The Maximum Clade Credibility (MCC) tree across all plausible trees was then computed from the BEAST trees using the TreeAnnotator program, with the first 10% of trees removed as burn-in.

3. Results

3.1 Phylogeny and major lineages of influenza A viral NP genes

Eight distinct lineages, denoted as A to H, were identified based upon the phylogeny of 7,517 NP sequences sampled between 1918 and 2010 (Figure 1). All the lineages are characterized by groupings with strong bootstrap support values. The seasonal human influenza lineage (A), the pandemic H1N1 2009 human lineage (B), and the classical swine lineage (C) are grouped together and appear to share a common ancestor, while the remaining five lineages are grouped together and are likely to have the same evolutionary origin (Figure 1). The annotations, representative sequences and main subtypes for each lineage and isolation periods of viruses within lineage are detailed in Table 1. Multiple subtypes were found in avian lineages, whereas only a few subtypes were found in non-avian lineages (Table 1).

Lineage A is composed of mostly human influenza viruses ($n = 2,647$), which includes H1N1 between 1918 and 1957, H2N2 between 1957 and 1968, H3N2 after 1968, and H1N1 after 1977, indicating that the NP genes of the viruses circulating in human populations were originally derived from the 1918 Spanish flu viruses (Figure 1). Lineage A also includes 38 viral strains from pigs and 1 from birds. Lineage B corresponds to pandemic H1N1 2009, which is grouped with lineage C, indicating the classical swine origin of the pandemic H1N1 2009 NP gene. Lineage C mainly consists of classical swine viruses ($n = 235$), but it also has 15 human viruses and 18 avian viruses. In addition, the triple-reassortant swine viruses, which contains genes from classic swine (NP, NS and M), human (HA, NA, PB1) and avian (PB2 and PA) influenza viruses, were also found in this lineage.

Three avian lineages (D, F and G) were identified corresponding to geographic distinctions (Figure 1). Lineage D is mainly composed of Eurasian avian viruses ($n = 1,805$), but with a few exceptions: 167 human, 116 swine, 8 environmental samples, 5 tiger, 2 mink, 3 feline, 1 stone marten and 3 canine viruses. Highly pathogenic H5N1 avian influenza viruses were found to be in lineage D as well. Lineage E is a Eurasian swine lineage derived from lineage D (Eurasian avian), possibly through direct transmission. Lineage F, although mostly from North American avian ($n = 1,090$), contain some strains from environmental samples (99), swine (4) and humans (2). Lineage G consists of Oceanian avian influenza viruses (35). Lineage H corresponds mainly to equine influenza viruses ($n = 92$), but a few are from canine (4) and swine (2). From the viewpoint of tree topology, a typical ladder-like tree structure was observed in equine viruses over time (Figure 1).

3.2 Substitution rates and times of most recent common ancestor (TMRCA) of influenza A NP lineages

The nucleotide substitution rates and TMRCA are summarized in Table 2. The estimated substitution rates of NP lineages ranged from 0.9×10^{-3} to 3.43×10^{-3} substitutions per site per year (nts/st/yr). Lineage B (pandemic H1N1 2009) was found to be the fastest evolving lineage, with a mean substitution rate of 3.43×10^{-3} and a 95% highest posterior density interval (HPD) of 2.7×10^{-3} - 4.12×10^{-3} . Lineage H (equine) is the slowest evolving lineage, with a mean estimate of 0.9×10^{-3} (95% HPD: 0.68×10^{-3} - 1.13×10^{-3}), which is significantly slower than those of other lineages.

The TMRCA of lineage A (human) was dated to 1906, and the 95% HPD interval was between 1884 and 1918 (Figure 2), which is 12 years before the outbreak of pandemic H1N1 1918. The mean TMRCA of lineage B (pandemic H1N1 2009) was estimated to be at Dec 17, 2008 (95% HPD: July 14, 2008 – Mar 30, 2009) (Supplementary Figure S1). The TMRCA of lineage C (classical swine) was estimated at 1911 (95% HPD: 1888 – 1927) (Supplementary Figure S2). The TMRCA of lineage D (Eurasian avian) was estimated to be 1920 (95% HPD: 1908 – 1927) (Supplementary Figure S3). The time of most recent common ancestor of lineage E (Eurasian swine) was estimated at 1972 (95% HPD: 1963 – 1978), seven years earlier than the first detection of Eurasian swine lineage in Belgium in 1979 (Supplementary Figure S4). The emergence of the North American avian lineage was dated to 1946 (95% HPD: 1932 – 1953) (Supplementary Figure S5). Finally, the most recent common ancestor of NP genes of equine influenza viruses can be traced back to 1947 (95% HPD: 1926 – 1959) (Supplementary Figure S6).

3.3 Selection pressure of NP lineages

The estimation of d_N/d_S ratios revealed different selection pressures occurred in different lineages (Table 3). The highest d_N/d_S ratio was observed in lineage H (equine: $d_N/d_S=0.13$). This is slightly higher than that of lineage A (human), which has a d_N/d_S ratio of 0.12. Interestingly, lineage G (Oceanian avian) and F (North American avian) appear to be under the strongest negative selection (d_N/d_S ratios: 0.04 and 0.05, respectively). The two swine lineages experienced similar selection pressures (0.09 for classical swine and 0.08 for Eurasian swine).

Positive and negative selection sites were identified in different NP lineages (Table 3, Figure 3). Positive selection sites were found in sequences of lineages A, B, C, D and E. The human lineage A was found to have 6 positive selection sites, whereas lineages B, C, D and E were observed to have only one or two sites. No positively selected sites were found in the equine lineage (H), Oceanian avian (E) or North American avian (H). The amino acid position 217 of NP of human lineage A and position 472 of NP of classical swine lineage C were found to be under positive selection. The position 472 is significant, in that structure analysis showed that it is located in an antibody binding region (Figure 3). In addition, the positive selection codon at position 217 is located in an alpha helix, which runs parallel to an adjacent antibody binding region, indicating possible strong viral adaptation to host immune system pressures. Similarly, the position 34 is located in a proposed CTL epitope site, while position 353 is located in the flanking region of a proposed CTL epitope where mutations are known to disrupt cytosolic processing of viral CTL epitopes.

4. Discussion

Lineage classification is critical for long-term influenza surveillance, because it offers a database of lineage information for detecting evolutionary changes of new viral strains (Chen et al., 2009). In this study, we conducted large-scale sequence analyses of the NP

segment (over 7,500 sequences) and identified eight distinct lineages. The results of our lineage classification and designation are largely in agreement with previous reports (see below); however, our study employs a much larger dataset and presents more detailed lineage annotations. The evolutionary relationship shown in our phylogenetic tree strongly suggests that human and classical swine lineages share the same evolutionary origin, which agrees with previous findings (Gorman et al., 1991). In addition, similar lineages (i.e., human, equine, classical swine, Eurasian avian, North American avian) were also designated based upon the evolutionary analyses of other gene segments (Chen et al., 2009; Dunham et al., 2009; Fourment et al., 2010; Gorman et al., 1990; Ito et al., 1991; Reid et al., 2004; Xu et al., 2011).

Lineage A includes human influenza viruses of three different subtypes isolated in different time periods. It is evident that the same NP gene remained in human influenza viruses, although two antigenic shifts occurred in 1957 and 1968, respectively. A similar situation was observed in the M gene (Furuse et al., 2009). Previous phylogenetic analyses revealed that the H2N2 virus, emerging in 1957, was a genetic reassortant between previously circulating human H1N1 virus dating from 1918 and avian viruses, with the novel H2, N2 and PB1 genes derived from Eurasian avian viruses (Kilbourne, 2006). In addition, the influenza A virus of the H3N2 subtype circulating in the human population since 1968 was also caused by reassortment between previously present human H2N2 virus and avian virus (H3, PB1) (Fang et al., 1981; Kawaoka et al., 1989).

With respect to evolutionary dynamics, the human lineage (A) of influenza NP has a substitution rate of 1.98×10^{-3} substitutions per site per year, a value close to previously reported ones (Jenkins et al., 2002). Although our mean TMRCA of the human lineage (A) is three years earlier than previously reported, the 95% HPD ranges are almost the same (1884–1918) (Smith et al., 2009a). A similar result was observed by Fourment et al., (2010) in the analysis of N1, with the most recent common ancestor dated back to 1906 (95% HPD: 1885–1905). Using the mutational bias towards an increased content of uracil in human influenza, Rabadan et al. (2006) estimated that the 20th century influenza viruses first diverged in 1910. Concerning the genetic origin of the 1918 pandemic influenza viruses, it appears that the NP gene, along with HA and NA, was first introduced into the human influenza viruses no more than two decades before that pandemic (Fourment et al., 2010).

The evolution of human influenza viruses is probably influenced by positive selection, gene flow, immune selection and competition (Bush et al., 1999; Grenfell et al., 2004). In contrast, immunological memory may not be substantially involved in the evolution of swine and domestic poultry influenza viruses because of their short life spans, which means influenza viruses may infect these non-human hosts only once in a lifetime (Meier-Ewert and Dimmock, 1970). Interactions with human immunity, partial cross-immunity, and competition between influenza viruses result in stronger immune selection at some amino acid sites among the influenza viruses infecting people (Bush et al., 1999). This was not seen in the avian influenza viruses. We found at most two positively selected sites and relatively low d_N/d_S ratios in avian lineages, supporting the view that no strong selection occurs in avian NP genes.

In the human lineage (A), the codons encoding residues 101 and 131 were found to be under positive selection in this study, which are potentially involved in the host adaptation (Forsberg and Christiansen, 2003). In an analysis of 246 NP sequences using the parsimony method, Suzuki (2006) also identified position 131 as a positive selection site. The codons at positions 31 and 34, which appeared to be under positive selection for Eurasian swine and Eurasian avian, respectively, are located in a region proposed to be involved in nuclear exportation (Neumann et al., 1997).

A novel swine-origin virus was detected in April 2009 in Mexico and spread rapidly to the rest of the world (Fraser et al., 2009). This pandemic (H1N1) 2009 influenza is believed to arise from a new reassortment between North American and Eurasian swine lineages (Christman et al., 2011; Garten et al., 2009; Smith et al., 2009b). The substitution rate for NP genes of pandemic (H1N1) 2009 estimated using sequences from the entire pandemic period in this study (3.43×10^{-3}) is higher than the rate estimated during the early outbreak period (2.59×10^{-3}) (Smith et al., 2009b). In addition, the d_N/d_S ratio of the pandemic 2009 NP (0.1) is obviously higher than the ratio of the closely related swine NP (0.05) (Smith et al., 2009b). The increase of the d_N/d_S ratio, which occurred after the inter-species transmission from swine to human, could be attributed to the adaptation to the new host and/or intensive epidemic surveillance (Smith et al., 2009b).

The evolution of equine lineage (H) seems to be quite interesting. First, a typical ladder-like tree structure was observed for the equine lineage from the phylogenetic tree (Figure 1). This phylogenetic pattern shows that viruses isolated at earlier times are located toward the root of the tree, while viruses isolated at a later time are presented farther from the root. Such tree topology reflects a process of temporal replacement of new viral strains and continual turnover of the viruses within this lineage due mainly to natural selection as previously suggested (Ferguson et al., 2003; Grenfell et al., 2004; Plotkin et al., 2002; Smith et al., 2004). Secondly, among all the eight lineages, the equine lineage was found to have the highest d_N/d_S ratio. The same d_N/d_S ratio (0.13) was reported by Murcia et al. (2011). In addition, the number of negative selection sites (24 using the SLAC method) was found to be much less than that found in other lineages, which might explain why the equine lineage has a higher d_N/d_S ratio. In this study, we did not find any positive selection site, but one positive selection site was identified by Murcia et al. (2011). This discrepancy might be attributed to the differences of significance levels and data sets used in the two studies. Concerning the substitution rate of equine viruses, a mean rate of 0.9×10^{-3} estimated in this study is consistent with previous estimates (Jenkins et al., 2002; Murcia et al., 2011). Compared with other hosts, however, equine viruses evolve much slower, reflecting host differences in their immune systems, population sizes, motilities and life spans. Finally, lineage H also includes canine influenza viruses, which provide evidence that equine influenza viruses have broken the species barrier and become established as a highly contagious respiratory pathogen of dogs (Crawford et al., 2005).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This publication was made possible by NIH grant numbers R01 LM009985-01A1. The authors also acknowledge the UCRCA, the University of Nebraska at Omaha, for continuous funding support to this research program. The authors are grateful to Andy Zhong and Todd Herpy for their help with structural analysis.

References

- Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 2010; 38(Suppl):W7–13. [PubMed: 20435676]
- 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; 82:596–601. [PubMed: 17942553]
- Bean WJ. Correlation of influenza A virus nucleoprotein genes with host species. *Virology.* 1984; 133:438–442. [PubMed: 6710867]

- Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. Predicting the evolution of human influenza A. *Science*. 1999; 286:1921–1925. [PubMed: 10583948]
- Chen JM, Sun YX, Chen JW, Liu S, Yu JM, Shen CJ, Sun XD, Peng D. Panorama phylogenetic diversity and distribution of type A influenza viruses based on their six internal gene sequences. *Virol J*. 2009; 6:137. [PubMed: 19737421]
- Chen R, Holmes EC. Avian influenza virus exhibits rapid evolutionary dynamics. *Mol Biol Evol*. 2006; 23:2336–2341. [PubMed: 16945980]
- Chen R, Holmes EC. Hitchhiking and the population genetic structure of avian influenza virus. *J Mol Evol*. 2010; 70:98–105. [PubMed: 20041240]
- Christman MC, Kedwaii A, Xu J, Donis RO, Lu G. Pandemic (H1N1) 2009 virus revisited: An evolutionary retrospective. *Infect Genet Evol*. 2011; 11:803–811. [PubMed: 21382522]
- Crawford PC, Dubovi EJ, Castleman WL, Stephenson I, Gibbs EP, Chen L, Smith C, Hill RC, Ferro P, Pompey J, Bright RA, Medina MJ, Johnson CM, Olsen CW, Cox NJ, Klimov AI, Katz JM, Donis RO. Transmission of equine influenza virus to dogs. *Science*. 2005; 310:482–485. [PubMed: 16186182]
- Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. *PLoS Biol*. 2006; 4:e88. [PubMed: 16683862]
- Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*. 2007; 7:214. [PubMed: 17996036]
- Dunham EJ, Dugan VG, Kaser EK, Perkins SE, Brown IH, Holmes EC, Taubenberger JK. Different evolutionary trajectories of European avian-like and classical swine H1N1 influenza A viruses. *J Virol*. 2009; 83:5485–5494. [PubMed: 19297491]
- Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 2004; 5:113. [PubMed: 15318951]
- Fang R, Jou W, Min, Huylebroeck D, Devos R, Fiers W. Complete structure of A/duck/Ukraine/63 influenza hemagglutinin gene: animal virus as progenitor of human H3 Hong Kong 1968 influenza hemagglutinin. *Cell*. 1981; 25:315–323. [PubMed: 6169439]
- Felsenstein J. Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Mol Biol Evol*. 2006; 23:691–700. [PubMed: 16364968]
- Ferguson NM, Galvani AP, Bush RM. Ecological and immunological determinants of influenza evolution. *Nature*. 2003; 422:428–433. [PubMed: 12660783]
- Forsberg R, Christiansen FB. A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. *Mol Biol Evol*. 2003; 20:1252–1259. [PubMed: 12777510]
- Fouchier RA, Munster V, Wallensten A, Bestebroer TM, Herfst S, Smith D, Rimmelzwaan GF, Olsen B, Osterhaus AD. Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. *J Virol*. 2005; 79:2814–2822. [PubMed: 15709000]
- Fourment M, Wood JT, Gibbs AJ, Gibbs MJ. Evolutionary dynamics of the N1 neuraminidases of the main lineages of influenza A viruses. *Mol Phylogenet Evol*. 2010; 56:526–535. [PubMed: 20434570]
- Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, Griffin J, Baggaley RF, Jenkins HE, Lyons EJ, Jombart T, Hinsley WR, Grassly NC, Balloux F, Ghani AC, Ferguson NM, Rambaut A, Pybus OG, Lopez-Gatell H, Alpuche-Aranda CM, Chapela IB, Zavala EP, Guevara DM, Checchi F, Garcia E, Hugonnet S, Roth C. Pandemic potential of a strain of influenza A (H1N1): early findings. *Science*. 2009; 324:1557–1561. [PubMed: 19433588]
- Furuse Y, Suzuki A, Kamigaki T, Oshitani H. Evolution of the M gene of the influenza A virus in different host species: large-scale sequence analysis. *Virol J*. 2009; 6:67. [PubMed: 19476650]
- Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, Sessions WM, Xu X, Skepner E, Deyde V, Okomo-Adhiambo M, Gubareva L, Barnes J, Smith CB, Emery SL, Hillman MJ, Rivaiier P, Smagala J, de Graaf M, Burke DF, Fouchier RA, Pappas C, Alpuche-Aranda CM, Lopez-Gatell H, Olivera H, Lopez I, Myers CA, Faix D, Blair PJ, Yu C, Keene KM, Dotson PD Jr, Boxrud D, Sambol AR, Abid SH, St George K, Bannerman T, Moore AL, Stringer DJ, Blevins P, Demmler-Harrison GJ, Ginsberg M, Kriner P, Waterman S, Smole S, Guevara HF, Belongia EA, Clark PA, Beatrice ST, Donis R, Katz J, Finelli L, Bridges CB, Shaw M, Jernigan DB, Uyeki TM, Smith DJ, Klimov AI, Cox NJ. Antigenic and genetic characteristics of swine-origin 2009

- A(H1N1) influenza viruses circulating in humans. *Science*. 2009; 325:197–201. [PubMed: 19465683]
- Gorman OT, Bean WJ, Kawaoka Y, Donatelli I, Guo YJ, Webster RG. Evolution of influenza A virus nucleoprotein genes: implications for the origins of H1N1 human and classical swine viruses. *J Virol*. 1991; 65:3704–3714. [PubMed: 2041090]
- Gorman OT, Bean WJ, Kawaoka Y, Webster RG. Evolution of the nucleoprotein gene of influenza A virus. *J Virol*. 1990; 64:1487–1497. [PubMed: 2319644]
- Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, Holmes EC. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*. 2004; 303:327–332. [PubMed: 14726583]
- Guan Y, Vijaykrishna D, Bahl J, Zhu H, Wang J, Smith GJ. The emergence of pandemic influenza viruses. *Protein Cell*. 2010; 1:9–13. [PubMed: 21203993]
- Holmes EC. Evolution in health and medicine Sackler colloquium: The comparative genomics of viral emergence. *Proc Natl Acad Sci U S A*. 2010; 107(Suppl 1):1742–1746. [PubMed: 19858482]
- Ito T, Gorman OT, Kawaoka Y, Bean WJ, Webster RG. Evolutionary analysis of the influenza A virus M gene with comparison of the M1 and M2 proteins. *J Virol*. 1991; 65:5491–5498. [PubMed: 1895397]
- Jenkins GM, Rambaut A, Pybus OG, Holmes EC. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol*. 2002; 54:156–165. [PubMed: 11821909]
- Kawaoka Y, Krauss S, Webster RG. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. *J Virol*. 1989; 63:4603–4608. [PubMed: 2795713]
- Kilbourne ED. Influenza pandemics of the 20th century. *Emerg Infect Dis*. 2006; 12:9–14. [PubMed: 16494710]
- Kuiken T, Holmes EC, McCauley J, Rimmelzwaan GF, Williams CS, Grenfell BT. Host species barriers to influenza virus infections. *Science*. 2006; 312:394–397. [PubMed: 16627737]
- Lu G, Rowley T, Garten R, Donis RO. FluGenome: a web tool for genotyping influenza A virus. *Nucleic Acids Res*. 2007; 35:W275–279. [PubMed: 17537820]
- Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefevre P. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics*. 2010; 26:2462–2463. [PubMed: 20798170]
- Meier-Ewert H, Dimmock NJ. Studies on antigenic variations of the haemagglutinin and neuraminidase of swine influenza virus isolates. *J Gen Virol*. 1970; 6:409–419. [PubMed: 5417271]
- Munster VJ, Baas C, Lexmond P, Waldenstrom J, Wallensten A, Fransson T, Rimmelzwaan GF, Beyer WE, Schutten M, Olsen B, Osterhaus AD, Fouchier RA. Spatial, temporal, and species variation in prevalence of influenza A viruses in wild migratory birds. *PLoS Pathog*. 2007; 3:e61. [PubMed: 17500589]
- Murcia PR, Wood JL, Holmes EC. Genome-Scale Evolution and Phylodynamics of Equine H3N8 Influenza A Virus. *J Virol*. 2011; 85:5312–5322. [PubMed: 21430049]
- Neumann G, Castrucci MR, Kawaoka Y. Nuclear import and export of influenza virus nucleoprotein. *J Virol*. 1997; 71:9690–9700. [PubMed: 9371635]
- Pensaert M, Ottis K, Vandeputte J, Kaplan MM, Bachmann PA. Evidence for the natural transmission of influenza A virus from wild ducts to swine and its potential importance for man. *Bull World Health Organ*. 1981; 59:75–78. [PubMed: 6973418]
- Plotkin JB, Dushoff J, Levin SA. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proc Natl Acad Sci U S A*. 2002; 99:6263–6268. [PubMed: 11972025]
- Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*. 2005; 21:676–679. [PubMed: 15509596]
- Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. *Bioinformatics*. 1998; 14:817–818. [PubMed: 9918953]
- Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet*. 2009; 10:540–550. [PubMed: 19564871]
- Rabadan R, Levine AJ, Robins H. Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. *J Virol*. 2006; 80:11887–11891. [PubMed: 16987977]

- Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. The genomic and epidemiological dynamics of human influenza A virus. *Nature*. 2008; 453:615–619. [PubMed: 18418375]
- Reid AH, Taubenberger JK, Fanning TG. Evidence of an absence: the genetic origins of the 1918 pandemic influenza virus. *Nat Rev Microbiol*. 2004; 2:909–914. [PubMed: 15494747]
- Ruigrok RW, Crepin T, Hart DJ, Cusack S. Towards an atomic resolution understanding of the influenza virus replication machinery. *Curr Opin Struct Biol*. 2010; 20:104–113. [PubMed: 20061134]
- Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA. Mapping the antigenic and genetic evolution of influenza virus. *Science*. 2004; 305:371–376. [PubMed: 15218094]
- Smith GJ, Bahl J, Vijaykrishna D, Zhang J, Poon LL, Chen H, Webster RG, Peiris JS, Guan Y. Dating the emergence of pandemic influenza viruses. *Proc Natl Acad Sci U S A*. 2009a; 106:11709–11712. [PubMed: 19597152]
- Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghvani J, Bhatt S, Peiris JS, Guan Y, Rambaut A. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*. 2009b; 459:1122–1125. [PubMed: 19516283]
- Snyder MH, Buckler-White AJ, London WT, Tierney EL, Murphy BR. The avian influenza virus nucleoprotein gene and a specific constellation of avian and human virus polymerase genes each specify attenuation of avian-human influenza A/Pintail/79 reassortant viruses for monkeys. *J Virol*. 1987; 61:2857–2863. [PubMed: 2441080]
- Stamatakis A, Ludwig T, Meier H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*. 2005; 21:456–463. [PubMed: 15608047]
- Suchard MA, Weiss RE, Sinsheimer JS. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol Biol Evol*. 2001; 18:1001–1013. [PubMed: 11371589]
- Suzuki Y. Natural selection on the influenza virus genome. *Mol Biol Evol*. 2006; 23:1902–1911. [PubMed: 16818477]
- Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology of influenza A viruses. *Microbiol Rev*. 1992; 56:152–179. [PubMed: 1579108]
- Xu J, Lu G. Evolution of influenza viral neuraminidase (NA) revealed by large-scale sequence analysis. *Influenza Other Respi Viruses*. 2011; 5(Suppl. 1):395–415.
- Ye Q, Krug RM, Tao YJ. The mechanism by which influenza A virus nucleoprotein forms oligomers and binds RNA. *Nature*. 2006; 444:1078–1082. [PubMed: 17151603]

Highlights

- We conducted large-scale phylogenetic analyses of influenza A NP genes
- 8 lineages identified: 3 host specific, 2 cross-host, 3 geographic-specific
- Substitution rates, TMRCA, and selection pressures vary in different lineages
- Evolution of influenza A viruses strongly associated with host adaptation

Figure 1.

The maximum-likelihood phylogenetic tree of influenza A viral NP genes. Eight lineages are denoted as A through H. The bootstrap support values are shown on the major nodes. The scale bars indicate the numbers of nucleotide substitutions per site.

Figure 2.
Maximum clade credibility (MCC) tree of the influenza A NP human lineage. The tree is scaled to the time (horizontal axis), with blue horizontal bars at nodes representing the 95% HPDs of TMRCA. The scale bar indicates the number of years before the present.

Figure 3. The structure of influenza A viral nucleoprotein (PDB: 2Q06), with positive selection sites denoted as yellow circles. Structural regions are denoted in different colors: pink for alpha-helices, red for beta sheets, and white for loops.

Table 1

Information of influenza A viral NP lineages and annotation

Lineage	Annotation	Isolation period	Representative sequence	Subtypes
A	Human	1918-2009	A/Brevig Mission/1/1918(H1N1)	H1N1, H2N2, H3N2
B	Pandemic H1N1 2009	2009-2010	A/Texas/05/2009(H1N1)	H1N1
C	Classical swine	1930-2009	A/swine/Iowa/15/1930(H1N1)	H1N1, H3N2, H1N2
D	Eurasian avian	1927-2009	A/fowl/Dobson/1927(H7N7)	H7N1, H2N3, H3N8, H3N3, H5N2, H5N1, H6N2, H9N2, H4N6, H7N3, H10N7
E	Eurasian swine	1979-2009	A/swine/Belgium/WVL1/1979(H1N1)	H9N2, H3N2, H1N2, H1N1, H5N1
F	North American avian	1953-2008	A/duck/Manitoba/1/1953(H10N7)	H5N2, H9N2, H6N2, H7N3, H3N8, H7N3, H2N2, H5N2, H1N1, H4N6, H12N5
G	Oceanian avian	1975-1984	A/shearwater/Australia/751/1975(H5N3)	H6N5, H3N8, H7N7, H11N3, H4N6, H4N4, H4N8, H6N2, H15N9, H1N2, H9N1, H6N1
H	Equine	1963-2008	A/equine/Miami/1/1963(H3N8)	H3N8

Table 2

Substitution rates and times of most recent common ancestor (TMRCA) of influenza A viral NP lineages*

Lineage	Substitution rate ($\times 10^{-3}$) subs/site/year				TMRCA (calendar year)			
	Mean	95% HPD lower	95% HPD upper		Mean	95% HPD lower	95% HPD upper	
A: Human	1.98	1.69	2.27		1906	1884	1918	
B: Pandemic (H1N1) 2009	3.43	2.7	4.12		17-Dec-08	14-Jul-08	30-Mar-09	
C: Classical Swine	2.29	1.99	2.58		1911	1888	1927	
D: Eurasian avian	3.41	2.83	4.01		1920	1908	1927	
E: Eurasian swine	2.8	2.4	3.21		1972	1963	1978	
F: North American avian	2.25	1.95	2.53		1946	1932	1953	
H: Equine	0.9	0.68	1.13		1947	1926	1959	

* Oceanian avian lineage (G): without enough sequences for reliable estimation of the parameters

Table 3

Selection sites detected using SLAC, FEL and IFEL methods (significant level = 0.05)

	Positively selected sites			No. of negatively selected sites			Mean d_N/d_S
	SLAC	FEL	IFEL	SLAC	FEL	IFEL	
A: Human	101, 131, 217, 451, 472	101, 131, 217, 451, 472	217, 450, 472, 473	328	352	322	0.12
B: pandemic H1N1 2009	None	None	353	66	155	20	0.1
C: Classical Swine	217	217	217, 472	267	318	279	0.09
D: Eurasian avian	34, 456	34, 456	456	416	418	417	0.08
E: Eurasian swine	None	None	31	180	264	179	0.08
F: North American avian	None	None	None	414	415	413	0.05
G: Oceanian avian	None	None	None	39	156	136	0.04
H: Equine	None	None	None	24	60	24	0.13