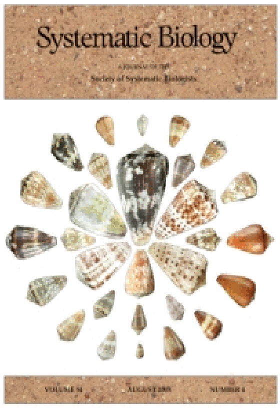


Systematic Biology

Genomic Analysis and Geographic Visualization of the Spread of Avian Influenza (H5N1)

To cite this Article: , 'Genomic Analysis and Geographic Visualization of the Spread of Avian Influenza (H5N1)', Systematic Biology, 56:2, 321 - 329
To link to this article: DOI: 10.1080/10635150701266848
URL: http://dx.doi.org/10.1080/10635150701266848

Syst. Biol. 56(2):321–329, 2007
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701266848

Genomic Analysis and Geographic Visualization of the Spread of Avian Influenza (H5N1)

Daniel Janies,[1] Andrew W. Hill,[2] Robert Guralnick,[2,3] Farhat Habib,[1,4]
Eric Waltari,[5] and Ward C. Wheeler[5]

[1] *Department of Biomedical Informatics, The Ohio State University, 3190 Graves Hall, 333 W. 10th Ave. Columbus, Ohio, 43210-1239, USA; E-mail: Daniel.Janies@osumc.edu*
[2] *Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado, 80309-0334, USA; E-mail: Robert.Guralnick@colorado.edu, Andrew.Hill@colorado.edu*
[3] *University of Colorado Museum, University of Colorado, Boulder, Colorado, USA, 80309-0265*
[4] *Department of Physics, The Ohio State University, 1040 Physics Research Building, 191 West Woodruff Avenue, Columbus, Ohio 43210-1117, USA; E-mail: farhat@pacific.mps.ohio-state.edu*
[5] *Division of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street, New York, New York, 10024-5193, USA; E-mail: ewaltari@amnh.org, wheeler@amnh.org*

Wild birds are known to carry all strains of influenza and, in theory, any of these strains could be the source of the next human pandemic. Most influenza infections in humans since 1968 have been attributed to influenza A, antigenic subtypes H3N2 or H1N1. However, due to recent and novel infections of humans with an avian strain of influenza (H5N1), a great deal of attention continues to be focused on the spread of H5N1. Key questions include which hosts have carried the virus from Asia to Russia, the Middle East, East and West Africa, and Europe, and which mutations of H5N1 have allowed it to be transmitted among various hosts, including humans. We have created an interactive genomic and geographic map using phylogenetic software and Google Earth (GE; earth.google.com) to reconstruct the evolution and spread of H5N1 influenza lineages over the past decade. Our results provide insight on competing hypotheses as to which avian hosts are responsible for the spread of H5N1. We show that the answers to these questions are temporally and geographically context specific. Various lineages of H5N1 spreading from southern China in the past decade have exploited diverse pathways in terms of avian host taxa and geography. By examining H5N1 phylogenies projected onto a globe and using character evolution to reconstruct host shifts, we studied, visually and statistically, whether key genotypes in viral proteins are correlated with the spread of the virus geographically and among hosts. We find that a key genotype (Lys-627 in polymerase basic protein 2, PB2) that allows for increased replication and virulence of H5N1 in laboratory mice (Subbarao et al., 1993; Shinya et al., 2004) is also significantly associated with mammalian hosts in the field (Table 1). In visualizations, this genotype appears to be prevalent in isolates of H5N1 circulating west of the East Asian–Australian flyway (Figure 1). However, this pattern is not supported statistically (Table 1). We also find that a genotype of a surface protein of H5N1 (Arg-110 in neuraminidase, NA) is correlated with geographically distinct clades but not strongly correlated with any host type.

In this paper, we demonstrate a workflow using a variety of computational tools to integrate diverse data into interactive genomic and geographic maps. These maps are suitable for analyses of the spread of disease agent genotypes that confer important phenotypes such as drug resistance or the ability to infect certain hosts. The data sets we examine include phylogenetic topologies, substitutions in sequence data, and time and place of isolation of disease agents and host species.

The Biology of Influenza Is Multifaceted

We use genomic and geographic maps to study the diversification of lineages of H5N1 influenza over the past decade. Wild aquatic birds (such as the order Anseriformes) have been implicated as the source of influenza viruses isolated from domestic birds (such as the order Galliformes) and mammals (orders Artiodactyla, Carnivora, Cetacea, and Primates; Webster et al., 1992). The current H5N1 outbreak originated in 1996 among anseriforms and spread to humans and chickens in Hong Kong in 1997 (Shortridge, 1999). Between 1997, when there were chicken culls in Hong Kong, and 2002, the most common hosts of H5N1 were anseriform birds and the virus was restricted to China. Since 2003, various lineages of H5N1 have spread throughout South East Asia, Russia, the Middle East, Europe, and Africa, using a wide variety of hosts. Many avian taxa (Charadriiformes, Accipitriformes, Corvidae, Ardeidae, Columbidae, and Passeriformes) as well as primate, carnivore, artiodactyl,

TABLE 1. The correlation between phenotypes and various genotypes calculated using Maddison's (1990) concentrated changes test. To correct for multiple testing we set the significance level at CCT ≤ 0.0125. Significant associations are in bold, and nearly significant ($0.0125 <$ CCT ≤ 0.05) associations are in italics.

Isolates in Dataset	Genotype	In or west of East Asian–Australian flyway	Isolated in 2005–2006 ?	Anseriform host?	Galliform host?	Mammalian host?
291	Isoleucine-99 in hemagglutinin	1.00	0.19	0.27	1.00	0.15
291	Asparagine-268 in hemagglutinin	*0.014*	0.061	0.105	1.00	1.00
291	Arginine-110 in neuraminidase	**0.007**	*0.035*	0.122	1.00	1.00
291	Lysine-627 in polymerase basic protein 2	1.00	1.00	1.00	0.48	<**0.00006**
351	Isoleucine-99 in hemagglutinin	0.073	1.00	0.258	0.193	0.170
351	Asparagine-268 in hemagglutinin	*0.042*	1.00	0.098	0.079	1.00
351	Arginine-110 in neuraminidase	0.034	1.00	0.084	0.0624	1.00
351	Lysine-627 in polymerase basic protein 2	0.184	1.00	*0.0164*	0.108	<**0.00001**

and arthropod hosts have been infected with H5N1 (Fig. 1a and aiTrees.kmz at www.systematicbiology.org).

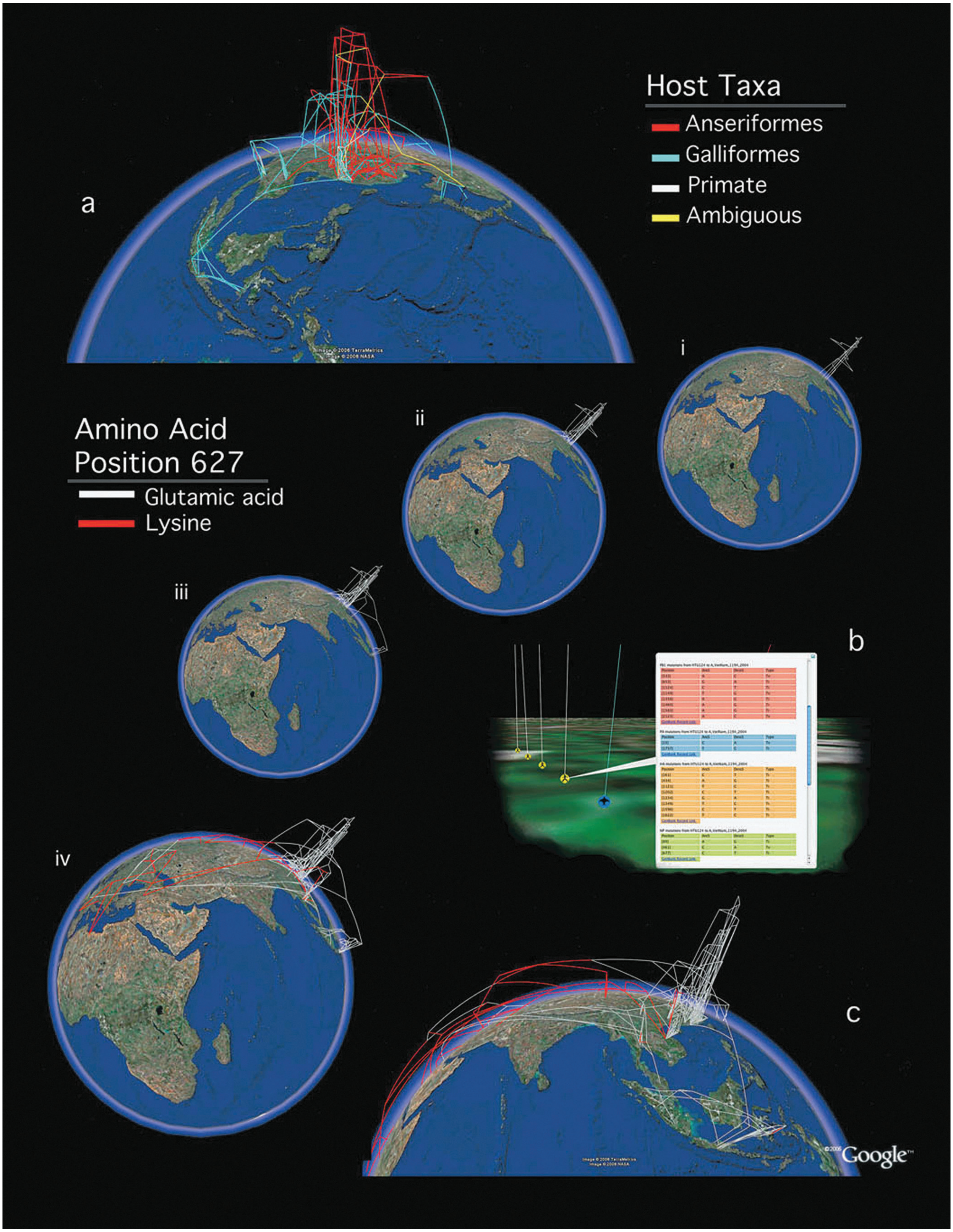
Direct human infection by avian strains of influenza A is considered rare (reviewed in Lipatov et al., 2004). However, the current outbreak of H5N1 influenza, thought to be strictly of avian origin (Li et al., 2004), has spread geographically and to novel hosts, including humans. Human H5N1 infections have been reported in Hong Kong in 1997–1999, in Vietnam, Cambodia, and Thailand in 2003–2006, in Indonesia in 2005–2007, in Laos and Nigeria in 2007, in China in 2003–2007, and in Azerbaijan, Turkey, Egypt, Iraq, and Djibouti in 2006 (WHO, 2007). As humans have little protective immunity to novel strains, continued transmission of avian strains to human populations and subsequent human-to-human transmission could have a devastating effect on public heath worldwide.

One evolutionary pathway for direct avian to human transmission requires mutations of the receptor-binding domains of the hemagglutinin (HA) protein of the influenza virus—especially those that effect changes in viral specificity (from NeuAcα2,3Gal sialo-oligosaccharides in avian cells to NeuAcα2,6Gal sialo-oligosaccharides in human cells; Stevens et al., 2006). Another key surface protein of influenza A that plays a role in host range is neuraminidase. Neuraminidase (NA) is a glycoprotein enzyme important to viral replication. The NA protein permits new viruses to escape host cells by removing terminal sialic acid from oligosaccharides bound to the HA protein. As a pair, the receptor binding (HA) and destroying (NA) proteins exhibit concerted evolution that mediates the range of hosts that can be infected (Kobasa et al., 1999). In relation to the recent outbreak of avian influenza, Chen et al., (2006) proposed that specific genotypes in HA and NA are associated with H5N1 from anseriform birds from lakes in central China. Here we used phylogenetic and geographic visualization of large data sets of genomes of H5N1 collected over the last decade to formulate and test specific hypotheses about the movement of key genotypes of H5N1 as it adapts to various hosts. Specifically, we used the concentrated changes test (CCT; Maddison, 1990) to evaluate whether putative key mutations in HA and NA (Stevens, et al., 2006; Chen et al., 2006) and PB2 (Subbarao et al., 1993) were associated with various bird taxa or mammals.

H5N1 Influenza Is a Concern of Public Health that Includes Several Important Social and Economic Issues

Avian influenza is not only complex and multidimensional in terms of biology but also raises several social and political issues. Pandemic influenza would have severe implications for public health, economic security, food safety, and wildlife conservation. For example, in

FIGURE 1. Screenshots from aiTrees.kmz found at systematicbiology.org. (a) Screenshot from an east (foreground) to west (background) perspective of a phylogenetic tree for 291 isolates. Branches of the tree are traced with color to represent the optimization of a character for taxonomic order of hosts. In this screenshot, only the hosts relevant for that point of view are provided in the key; greater detail can be found in aiTrees.kmz. (Images i to iv) A temporal series of screenshots showing the movement of the genotype Lys-627 in PB2, which increases the ability of the virus to replicate in mammalian hosts, as red branches on a phylogenetic tree for 351 isolates. Images i to iv show movement of Lys-627 in PB2 as of 2000, 2002, 2004, and 2006, respectively. (b) A close-up view of a description box containing an account of mutations for that isolate (see detail for all isolates and HTUs aiTrees.kmz). (c) A south (foreground) to north (background) view of avian influenza spread from East Asia, showing Lys-627 position in PB2 character optimization as colored branches. Earth background image sources: Google, TerraMetrics, and NASA.



PB1 mutations from HTU124 to A,VietNam,1194,2004

Position	Anc5	Desc5	Type
[illegible]	A	C	Tv
[illegible]	G	A	Ts
[illegible]	C	T	Ts
[illegible]	T	G	Tv
[illegible]	A	C	[illegible]
[illegible]	A	G	Ts
[illegible]	A	G	Ts
[illegible]	A	C	Tv

GenBank Record Link

PA mutations from HTU124 to A,VietNam,1194,2004

Position	Anc5	Desc5	Type
[illegible]	C	A	Tv
[illegible]	T	C	Ts

GenBank Record Link

HA mutations from HTU124 to A,VietNam,1194,2004

Position	Anc5	Desc5	Type
[illegible]	C	T	Ts
[illegible]	A	G	Ts
[illegible]	T	C	Ts
[illegible]	C	T	Ts
[illegible]	G	A	Ts
[illegible]	T	C	Ts
[illegible]	C	T	Ts
[illegible]	T	C	Ts

GenBank Record Link

NP mutations from HTU124 to A,VietNam,1194,2004

Position	Anc5	Desc5	Type
[illegible]	A	G	Ts
[illegible]	C	A	Tv
[illegible]	C	T	Ts

GenBank Record Link

the United States alone it is projected that 15% to 35% of the human population would be affected and the costs could range from $71.6 to $166.5 billion (Gerberding, 2005). Recent papers have pointed out gaps in our knowledge of the patterns of influenza transmission among domestic or wild birds (Olsen et al., 2006; Ducatez et al., 2006). Some advocate that wild birds spread H5N1 over long distances, whereas others contend that the globalization of trade in poultry and wildlife are responsible for the spread of H5N1 (Van Borm et al., 2005; Karesh et al., 2005; Chen et al., 2006; Kilpatrick et al., 2006). The debate among these groups is reflective of the interests of various stakeholders (Normile, 2006a).

Rapidly addressing questions about the spread of H5N1 requires that all data are made available as quickly as possible to the global research community (Salzberg et al., 2006). However, there is a great deal of H5N1 sequence data in private hands (Enserink, 2006). Recently, a central database to share avian flu data after publication has been proposed (Bogner et al., 2006; www.gisaid.org) but there remain sociological impediments to data sharing. The underlying causes for the privacy of data include the need for researchers to receive credit for data and primary publications and the desire of governments and individuals to protect their economic interests. For example, subsistence duck and chicken farmers and vendors of live birds in markets may be unwilling to report H5N1 outbreaks in their flocks if they are not guaranteed compensation for culled birds (Shortridge, 1999). Thus far, the losses in the poultry sector in South East Asian countries that have had H5N1 epidemics have ranged between 1% and 2% of GDP due to culls and banned exports (Brahmbhatt, 2005). Nongovernmental organizations have their own interests and concerns regarding H5N1 avian influenza. Advocates for the conservation of wildlife fear that wild birds could be treated as primary or "scapegoat" carriers of H5N1 (Blythman, 2006) and subject to eradication programs.

Given the biological and social complexity of the problem of avian influenza, it is important that diverse data be synthesized in a form communicable to a wide range of stakeholders. Our aim is to show the utility of phylogeographic visualization when used in concert with more traditional character evolution approaches. These visualizations can be used to synthesize diverse data on pathways of avian influenza spread and communicate these results to a wide audience. To these ends, we combined large-scale phylogenetic analysis of H5N1 genomes with novel visualization and mapping techniques using GE. We then studied geographic and temporal patterns of putative key mutations and host use of H5N1 (Fig. 1 and aiTrees.kmz at www.systematicbiology.org).

Genome and Isolate Sampling

To examine avian influenza evolution, we performed a phylogenetic analysis of the H5N1 genome from 291 isolates, 259 of which were complete genomes. After initial submission of our work to *Systematic Biology*, many new H5N1 genomes from Vietnam, Europe, the Middle East, and Africa were released. These new data are the result of international collaborations catalyzed via the Influenza Genome Sequencing Project (msc.tigr.org/infl_a_virus/index.shtml) and were released to GenBank (www.ncbi.nlm.nih.gov), prepublication, soon after final assembly. To enhance our sampling of H5N1 genomes from Europe, the Middle East, and Africa, we added the new genomes and removed some partial genomes, to create a second data set of 351 complete genomes. Tables of GenBank accession numbers for the 291 and 351 isolate data sets are provided as supplemental tables (suptable1.xls and suptable2.xls at www.systematicbiology.org). A third data set of 80 isolates was created to accommodate the recently released sequences of Smith et al., (2006a) of Chinese isolates that bear close relationship to the clade of H5N1 that recently moved west of the East Asian–Australian flyway.

Multiple Alignments

Multiple-sequence alignment of nucleotide and amino acid data was performed with MUSCLE (Edgar, 2004) under default parameters. A small number of internal gaps were introduced in nucleotide sequences for HA, PB2, PB1, PA, NA, and NS segments by multiple alignment. The alignments of segments NP and M lacked internal gaps. Leading and trailing gaps were not considered in tree search but all nucleotide positions and internal gaps were considered.

Tree Search, Character Coding, and Optimization

Phylogenetic analyses were conducted using concatenated alignments of eight nucleotide segments. The tree search strategy used for each data set included 1000 replicates of a heuristic tree search in TNT (Goloboff et al., 2005) under the parsimony criterion. Edit costs for transitions, transversions, and insertion-deletion events were equally weighted. As it is likely a good representative of the isolates in circulation at the origins of the H5N1 epidemic and is represented by a full genomic sequence, the strain A/Goose/Guangdong/1/96 was used as the outgroup in the analyses of 291 and 351 isolates. The TNT command "xmult = lev5" was used to specify a tree search that consisted of a Wagner tree with a random addition sequence followed by heuristic refinement procedures: TBR branch swapping, sectorial searches, drifting, and tree fusing. When refinement was complete, 113 trees of presumed minimum length (15945 steps) were saved for the 291-isolate data set. The 291-isolate data set was also subjected to 500 bootstrap resampling pseudoreplicates. The full strict consensus and majority-rule consensus tree with bootstrap values are too large for the printed page; thus, scalable documents in portable document format (.pdf) are provided as supplemental data (supfig1.pdf and supfig2.pdf at www.systematicbiology.org).

The alignment and phylogenetic methods used to analyze the 351-isolate data set were identical to that described for the 291-isolate data set. For the 351-isolate data set, 635 trees of presumed minimum length (18034 steps) were found. The strict consensus and 50% majority-rule consensus tree with bootstrap support values for the 351-isolate data set are provided as supplemental data (supfig3.pdf and supfig4.pdf at www.systematicbiology.org).

In the 80-isolate data set, we included isolates from west of the East Asian–Australian flyway, the isolates (A/migratory duck/Jiangxi/1653/) from China that were the outgroup to the isolates from the west (referred to as the "west clade" below), and the relevant new sequences of Smith et al. (2006a). The sequences of Smith et al. (2006a) relevant to the west clade were identified by adding Smith's isolates to the larger analyses and then subsampling only the isolates that fell within or sister to the west clade. The tree search of the 80-isolate data set produced 391 trees of presumed minimum length (1692 steps). The strict consensus and 50% majority-rule consensus tree with bootstrap support values for the 80-isolate data set are provided as supplemental data (supfig5.pdf and supfig6.pdf at www.systematicbiology.org).

For the 291-isolate data set, for each of the eight genome segments, a complete accounting of mutations optimized to ancestor-descendent branches implied by the tree (apomorphy list) was calculated using the commands "-diagnose" in POY (Wheeler, 1996; Wheeler et al., 2005). The tree used in these calculations was a binary representation of a minimum length tree randomly chosen using the TNT (Goloboff et al., 2005) command "randtrees" from the pool of minimum length trees. The list of nucleotide mutations optimized to each branch for each locus was converted into a Keyhole markup language (KML) file suitable for viewing with GE (aiTrees.kmz at www.systematicbiology.org). The minimum length tree and the diagnose command was also used to optimize character data such as host for the 291-isolate data set or amino acid genotype of the 351-isolate data set. The resulting character optimizations and apomorphy lists were used to code branches with various colors and fill description boxes used in the visualizations (aiTrees.kmz at systematicbiology.org).

Visualization

The desktop mapping client GE is freely available, can run on multiple platforms, and is user-friendly. The GE interface allows for interactive visualization across geographic and temporal scales for the entire globe. The Network Link function in GE's KML was used to visualize phylogenetic subtrees containing H5N1 isolates from each year during the period 1996–2006. This sequence of subtrees is useful for exploration of temporal and regional spread of H5N1. In addition, subtree layers can be used in conjunction with existing layers in GE such as borders, populated places, and geographic features, or user-supplied graphics for further study of the spread of H5N1 in a variety of contexts.

Apomorphy lists for all influenza genome segments, georeferenced strain locations, and phenotypic and other metadata (e.g., host taxa, date and place of isolation) were aggregated into a KML file for upload into GE. Once the KML file is loaded in the GE application, the user may reorder, rename, or remove different layers, rotate and zoom the point of view of the earth and phylogeny, select terminal or ancestral nodes in which to view mutations in description windows (Fig. 1 and aiTrees.kmz at www.systematicbiology.org), and save reconstructions of interest to their local disk or printer.

To construct the KML file of the three-dimensional H5N1 phylogeny, each terminal taxon and hypothetical taxonomic unit (HTU) was assigned a set of coordinates in three-dimensional space (latitude, longitude, and altitude). Using locality data from Genbank for each viral genome and an online gazetteer (www.biogeomancer.org or www.getty.edu/research/tools/vocabulary/tgn/), we determined the latitude and longitude for each terminal taxon in our tree. Our ability to know the exact location of isolation was often constrained by the quality of the metadata. For example, a locality description such as "Hong Kong" was assigned coordinates for the centroid of Hong Kong.

The internal nodes or HTUs were assigned latitude and longitude coordinates based on the geographic location of their descendants. First we calculated the centroid of all descendants within the clade subtended by an HTU to assign longitude and latitude to that HTU. All terminal taxa were given an altitude of zero, thus anchoring the tree to the ground (Fig. 2). The altitude of HTUs was assigned based on the hierarchical position from the tip of the phylogenetic tree—the root node is assigned the highest altitude; less inclusive HTUs are assigned lower altitudes. For example, any HTU with only terminal taxa as descendants is given an altitude of a. Any HTU ancestral to other HTUs was given an additional altitude of b. Thus for any HTU, its altitude $= a + [(n - 1) \times b]$, where n is the number of HTUs from the tip of the tree to the node of interest. In our visualizations, we chose 198,000 meters for the variable a, determining the minimum altitude of any nonterminal taxon. We chose 66,000 m for the variable b, which was assigned in quanta representing the vertical expansion of steps in the hierarchy of ancestral nodes. These values of a and b were selected because

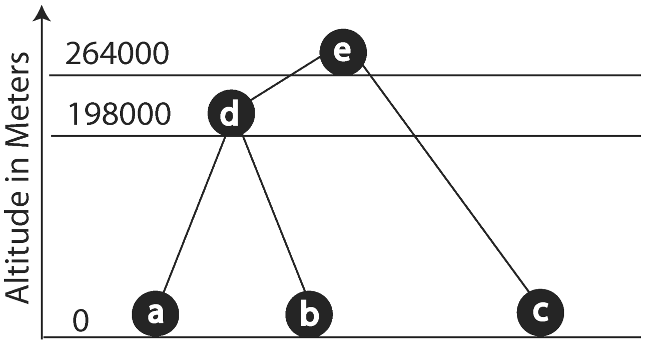


FIGURE 2. A phylogenetic tree, where a, b, and c are terminal taxa and are located at ground level on the surface of the globe. The internal nodes, or hypothetical taxonomic units (HTUs), d and e are elevated above the ground and are connected to their descendants by the black branches. HTU d is given an altitude of 198,000 m, because it has entirely terminal descendants. HTU e is ancestor to HTU d, so HTU e is given an altitude of the cumulative altitude of its descendants (HTU d at 198,000 m) plus 66,000 m, or 264,000 m.

they provided the most visually clear configuration for our trees. We are investigating the use of altitude variables to represent quantitative data, such as phylogenetic branch length.

In early attempts at tree projection, we encountered a problem visualizing overlapping clades. For example, 90 strains had a location of "Hong Kong." If street level locality data were available, this visualization would be feasible and accurate. However, given the coarse resolution of locality data often provided in a Genbank record, we needed a procedure to evenly distribute each terminal within the boundaries of the location provided by the Genbank record. For this, we assigned the first terminal found within the apomorphy list the centroid latitude and longitude for its location. Then each terminal with an identical location was moved $y \times s$ degrees in longitude away from its actual longitude, where s is the number of points with the same starting coordinates and y is the number of degrees to move the point (let $y = 0.001$ degrees of longitude in this study). This procedure created an east to west row of all isolates occurring at the same locality (Fig. 1b). One could also vary the latitude position for overlapping georeferences.

Once coordinate data for all terminal and HTU nodes were satisfactorily calculated, tree branches between corresponding nodes were drawn using GE's Pathway function. For each ancestor-descendant relationship in the original tree, we drew a line between the two points connecting the latitude, longitude, and altitude coordinates of each point into a directed acyclic graph. Tree branch lines were coded with colors to represent various character states (e.g., genotype or host) reconstructed in the phylogenetic analysis (see Fig. 1 and aiTrees.kmz at www.systematicbiology.org). Terminal nodes were given placeholders using the Placemark function of GE. This function allows for simple iconic representation of character state data for isolates. We created a simple library of icons that allow the user to determine the host type from which the strain was isolated.

We used popup windows in GE, termed description boxes, to contain nucleotide and amino acid character states as defined in the apomorphy lists (Fig. 1b). The description boxes allow the user to interact with the tree and apomorphy data to explore key mutations associated with host shifts or markers that can be used to genotype viral isolates in an area of interest. These windows also include external links to public data sources such as GenBank. We took the year of the sequence and added the TimeSpan KML function to animate the movement of the virus over the past decade. Example KML files are available as supplemental data (aiTrees.kmz at www.systematicbiology.org).

Statistical Validation of Visualized Patterns

We use a statistical test to examine hypotheses generated by the visualization. The concentrated changes test (CCT; Maddison, 1990) calculates the probability of a dependent character (such as host) having changes mapped to the same branches of a phylogeny as changes in independent character (such as genotype) by chance, assuming a null model where changes, in either character are randomly distributed. If changes in each of the two characters occur on several of the same branches of the phylogeny, the CCT value suggests significant (≤ 0.05) character correlation. If the changes in each character of interest share few or no common branches, then the CCT value suggests no significant (>0.05) relationship among the characters. We corrected the significance level for multiple testing to ≤ 0.0125 by dividing 0.05 by the four tests per phenotype per data set. We calculated the CCT using MacClade (Maddison and Maddison, 2003) with exact calculations. In case of ambiguous ancestral reconstruction, we use the DELTRAN option to resolve the ambiguity.

Results

Geographic Spread of H5N1 and Associated Changes in Host Use

Early in the epidemic (1996–2002), several lineages of H5N1 were carried by anseriform and galliform birds throughout Hong Kong and Southern China (Fig. 1a and aiTrees.kmz). Subsequently in 2001–2006 diverse avian hosts from the taxa Charadriiformes, Accipitriformes, Corvidae, Ardeidae, and Columbidae were infected in China. In 2001–2003, swine were infected with H5N1 in China. Following the outbreak in Hong Kong in 1997, human cases were next reported by Chinese officials in 2003 (WHO, 2006).

In 2003 and 2004, a single lineage of H5N1 entered Korea and Japan from China. This lineage then radiated in galliform, corvid, and arthropod hosts within Japan and Korea. In 2005, South Korea was considered free of H5N1. In late 2006, there are reports of a novel outbreak in Korea (OIE, 2006). It is not known at this time if the viruses that underlie the 2006 outbreak in Korea are related to the original invading lineages or a novel lineage of H5N1 (Normille, 2006b).

Starting in 2003 among galliforms from southern China, three distinct lineages of H5N1 formed in Asia. One lineage circulates exclusively in Vietnam, another exclusively in Indonesia, and a third can be found in Vietnam, Thailand, Malaysia, Laos, and Cambodia. In 2004, each lineage radiated regionally via galliforms, and anseriforms (Chen et al., 2006; Smith et al., 2006b; Boltz, 2006). Thus far, the clade that is exclusively in Vietnam does not affect human or mammalian hosts. However, H5N1 isolates from swine in Fujian region of China are members of a sister group to the lineage exclusive to Vietnam. The lineage that spans much of Southeast Asia infects mammals including humans. The Indonesia clade infects humans, felids, and a variety of birds.

In 2004, after H5N1 began its radiation in Southeast Asia but before its arrival in Europe, an isolate of the Thailand strain of H5N1 was carried by eagles smuggled from Thailand to Belgium (Van Borm et al., 2005). However, even without knowing this information, the case presents a clear anomaly in our Google Earth–based phylogeographic visualization, in which a single branch makes an enormous geographic leap (aiTrees.kmz in supplemental data). Such anomalies provide valuable insights into the spread of H5N1 and when significant can be used to prompt further investigation into potential causes.

Although the smuggled eagles were seized before H5N1 could infect other hosts, the event represented a potential pathway for spread of H5N1 to Europe distinct from the radiation of H5N1 seen in 2005–2006. In 2005–2006, several lineages of H5N1 moved west of 100 degrees east longitude and out of the East Asian–Australian bird flyway. These lineages originated in anseriform birds from China and subsequently infected galliform birds in China and Russia. In late 2005 and early 2006, these Russian and Chinese lineages of H5N1 persisted in those regions and expanded their range to diverse birds and mammals (including humans) in Mongolia, the Middle East, Africa, and Europe. A recent paper (Smith et al., 2006a), released 556 HA and PB2 sequences from viruses isolated from various birds in China in 2006. Based on this partial data, some of these isolates (A/Guinea foul/Shantou/1341/2006) are closely related to isolates at the origin of the clade that moved west of 100 east longitude (supfig5.pdf). Based on our preliminary results, we note that there is not only westward movement of this clade (Ducatez et al., 2006; Webster and Govorkova, 2006) but eastward movement as well—from Russia back to China and Mongolia (supfig5.pdf and aiTrees.kmz at www.systematicbiology.org). Full genome data from these regions of the world will be of use for resolving the biogeography and host utilization of this widely distributed and actively transported H5N1 clade.

Genotypes Associated with Various Hosts

We see a strongly supported association between the genotype Lys-627 in PB2 and mammalian hosts in the 291- and 351-isolate data sets (Table 1). This genotype does not occur exclusively in mammals but is of interest because it is experimentally associated with increased replication and virulence of the H5N1 virus in laboratory mice (Subbarao et al., 1993; Shinya et al., 2004). In the 351-isolate data set the association between Lys-627 in PB2 and anseriform hosts is marginally nonsignificant under the conservative (CCT $\leq$ 0.0125) significance level that we have set.

We see no genotype that is significantly associated with certain host types in the amino acid positions of the surface proteins (HA and NA) examined in the 291- or 351-isolate data sets. Mutations of HA in amino acid positions 226 and 228, which mediate a shift from avian to human specificity in seasonal influenza strains of subtype H3 (Stevens et al., 2006), are virtually invariant at Gln-226 and Gly-228 among the 291 and 351 isolates of H5N1 that we considered. Although Arg-110 in NA was proposed as a signature for H5N1 adaptation to migratory waterfowl (Chen et al., 2006), this genotype is not significantly correlated with any particular host (Table 1).

Spread of Various Genotypes over Time and Space

In genotypes of HA amino acid positions 99 and 268, we see virtually no variation from the Ala-99 Tyr-268 genotype within the East Asian–Australian flyway in the 291- and 351-isolate data sets, except that a few isolates have Thr or Val at position 99. To the west of East Asian–Australian flyway, however, the HA genotype Ile-99 Asn-268 is prevalent, with the sole exception of a single branch of the tree representing isolates of H5N1 from eagles smuggled from Thailand to Belgium in 2004 (Van Borm et al., 2005). The bias of Asn-268 in the west is statistically significant at the CCT $\leq$ 0.05 level but is marginally nonsignificant at the CCT $\leq$ 0.0125 level (Table 1). We found nonsignificant correlation in the 291- or 351-isolate data sets between HA amino acid positions 99 and 268 and dependent characters of time, anseriform host, galliform host, and mammalian or avian host (Table 1).

Genotype Arg-110 of the surface protein NA is significantly correlated with viruses isolated west of the East Asian–Australian flyway for at least the 291-isolate data set (CCT $\leq$ 0.0125 in 291-isolate data set but CCT $\leq$ 0.05 in 351-isolate data set; Table 1). The correlation of the genotype Arg-110 of NA with isolates west of the East Asian–Australian flyway is nearly significant for the 291-isolate data set but nonsignificant for the 351-isolate data set.

Despite the visual appeal of a potential correlation (Fig. 1), the CCT does not indicate a strong correlation between Lys-627 in PB2 and the 2005–2006 date of isolation or in viruses isolated west of the East Asian–Australian flyway. Avian influenza data are updated frequently and the phylogeny, visualization, and statistical tests can be rerun accordingly.

Discussion

We used both visual and statistical approaches to examine two important questions: (1) Can phylogeographic visualization provide insight as to which bird taxa spread H5N1 (Chen, et al., 2005, 2006; Van Borm et al., 2005; Melville and Shortridge, 2006; Kilpatrick et al., 2006)? (2) Do we see distinct temporal and spatial patterns in putatively key mutations in H5N1's proteins? The mutations we chose to track are thought to be important to infection and replication of H5N1 in various hosts such as mammals, anseriform, or galliform birds (Subbarao, 1993; Chen et al., 2006; Stevens et al., 2006).

Which Avian Taxa Spread H5N1?

Our results suggest that the avian taxa responsible for the spread of H5N1 are context-specific temporally and geographically. Recent work has shown a tendency to pose the question of H5N1 spread in the terms of whether "migratory or nonmigratory" birds "traded legally or illegally" are responsible for movement of the disease (Van Borm et al., 2005; Karesh et al., 2005; Chen et al., 2006; Normile, 2006). However, the metadata associated with public genomic data often lack annotations on bird behavior or husbandry. In some cases, the metadata associated with a genomic record do include trade information and from those cases we can be certain that H5N1 has entered different regions of the world via different host pathways. For example, the case of H5N1-infected Thai eagles seized in Belgium in 2004 is clearly illegal transcontinental transport (Van Borm et al., 2005). However, other putative cases where illegal trade plays a role in the spread of H5N1 are difficult to determine. Records and maps of illegal trade are very difficult to produce due to the nature of the underlying data (Rosenthal, 2006). Recently, Kilpatrick et al., (2006) have made an attempt to combine trade data from the Food and Agricultural Organization (FAO) of the United Nations and a small sample of sequence data (∼12.8% of the genome) from the HA gene in order to predict further spread of H5N1. However, directly linking the sequences from Genbank to the FAO data is inexact at best and does not solve the problem of poor metadata associated with the genomic data.

Despite rapid genomic sequencing, improved understanding of virus–host cell receptor interactions, and forecasts of the movement H5N1, the issue of the provenance of various forms of data needs to be addressed if we are to make stronger inferences about the global impact of this disease. Our phylogeographic visualizations demonstrate that it is possible to reconstruct specific genomic changes in the virus in a geospatial context, but we are still limited by the quality of the metadata available with genomes. Thus, just as there has been a plea for more public data (Enserink et al., 2006; Salzberg et al., 2006), surveillance and eradication programs for infectious diseases will benefit greatly from integrating diverse and accurate sampling metadata with genomic data. Metadata of interest include species identifications rather than common names for hosts, precise georeferences and dates for viral isolation, data on the community structure or husbandry of host species, and data from legal trade and seizures of illegally traded host species (Gilbert et al., 2006; Vam Borm et al., 2005).

Have Key Genotypes that Underlie Host Switching Been Identified among Lineages of Avian Influenza (H5N1)?

The genotypes for surface proteins (HA and NA) that we examined are either invariant or have distinct geographic and temporal patterns in their variation. These genotypes in surface proteins do not correlate statistically or visually to host use by H5N1 as determined by our character mapping approaches, despite the experimentally indicated roles in mediating the efficacy and specificity of viral surface proteins with receptors on various host cells (Stevens, 2006; Chen, 2006). In contrast, we do see strong correlation between Lys-627 in PB2 and mammalian hosts. This genotype is not exclusive to H5N1 in mammals; it also occurs in H5N1 lineages circulating in birds. In summary, suggestions of key genotypes for the spread of H5N1 to various hosts based on experimental mutagenesis represent a prognostic inference on what mutations we should be tracking. Our study of the actual geographic variation of H5N1 mutations and host shifts puts these experimental inferences in a real world context and checks these prognoses against the data derived from isolates actually circulating in the field.

Multidimensional Analysis and Effective Communication of Results on Large Complex Systems Are Important

As demonstrated by large-scale sequencing projects (Ghedin et al., 2005; Obenauer et al., 2006), infectious diseases are commonly studied using comparative genomics. However, our ability to analyze, visualize, and interpret large genomic data sets in phylogenetic, geographic, and societal contexts is nascent. The spread of infectious disease is one of the most important issues facing faunas, the food industry, world economy, and governments. Although there is significant public, research, and media interest in genomic and geographic analyses of H5N1, these results are complex and difficult to communicate to a wide audience. As Google Earth is freely available, any user can open the keyhole mark-up files (aiTrees.kmz provided as supplemental data) to view H5N1 lineage diversification by region, host type, genotype, or time of interest. We have developed a general workflow that can address both problems of analysis and communication. As more data on avian influenza and other emerging infectious diseases become available, analyses and visualizations can be readily updated to track specific mutations over time, hosts, and geography.

Acknowledgements

We acknowledge this material is based upon work supported by, or in part by, the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-05-1-0271. We thank Lisa Gugenheim, Merrily Sterns, and Tanya Das of the American Museum of Natural History, for their help in management of the project. FH, AH, and DJ acknowledge the support of the Department of Biomedical Informatics and the Ohio State University Medical Center. The viral isolates and much of the sequence data were produced by many individuals and we thank them for sharing their data via the National Institutes of Health's Genbank. Genomic sequencing of key Avian Influenza isolates was funded by the National Institutes of Health, National Institute of Allergic and Infectious diseases, and performed at the Institute for Genomic Research. We thank Rod Page, David Kidd, and Deborah Ciszek for their help in the reviewing and editorial process.

References

Blythman, J. 2006. So who's really to blame for bird flu? The Guardian. June 7. www.guardian.co.uk/birdflu/story/0,,1791954,00.html

Bogner, P., I. Capua, N. Cox, D. Lipman, et al., 2006. A global initiative on sharing avian flu data. Nature 442:981.

Boltz, D. A., B. Douangngeun, S. Sinthasak, P. Phommachanh, S. Rolston, H. Chen, Y. Guan, J. Peiris, G. Smith, and R. Webster. 2006. H5N1 influenza viruses in Lao People's Democratic Republic. Emerg. Infect. Dis. 12:1593–1595.

Brahmbhatt, M. 2005. Avian influenza: Economic and social impacts. web.worldbank.org/WBSITE/EXTERNAL/NEWS/0,,contentMDK:20663668∼pagePK:34370∼piPK:42770∼theSitePK:4607,00.html

Chen, H., G. Smith, K. Li, J. Wang, X. Fan, J. Rayner, D. Vijaykrishna, J. Zhang, L. Zhang, C. Guo, C. Cheung, K. Xu, L. Duan, K. Huang, K. Qin, Y. Leung, W. Wu, H. Lu, Y. Chen, N. Xia, T. Naipospos, K. Yuen, S. Hassan, S. Bahri, T. Nguyen, R. Webster, J. Peiris, and Y. Guan. 2006. Establishment of multiple sublineages of H5N1 influenza virus in

Asia: Implications for pandemic control. Proc. Natl. Acad. Sci. USA 103:2845–2850.

Chen, H., G. Smith, S. Zhang, K. Qin, J. Wang, K. Li, R. Webster, J. Peiris, and Y. Guan. 2005. Avian flu H5N1 virus outbreak in migratory waterfowl. Nature 436:191–192.

Ducatez, M., C. Olinger, A. Owoade, De Landtsheer S., W. Ammerlaan, H. Niesters, A. D. Osterhaus, R. Fouchier, and C. Muller. 2006. Avian Flu: Multiple introductions of H5N1 in Nigeria. Nature 442:37.

Edgar, R. C. 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5: 113.

Enserink, M. 2006. Avian influenza: As H5N1 keeps spreading, a call to release more data. Science 311:1224.

Gerberding, J. 2005. Pandemic planning and preparedness. www.cdc.gov/Washington/testimony/in05262005.htm

Ghedin, E., N. Sengamalay, M. Shumway, J. Zaborsky, T. Feldblyum, V. Subbu, D. Spiro, J. Sitz, H. Koo, P. Bolotov, D. Dernovoy, T. Tatusova, Y. Bao, K. St. George, J. Taylor, D. Lipman, C. Fraser, J. Taubenberger, and S. Salzberg. 2005. Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. Nature 437:1162–1166.

Gilbert, M., P. Chaitaweesub, T. Parakamawongsa, S. Premashthira, T. Tiensin, W. Kalpravidh. H. Wagner, and J. Slingenberghet. 2006. Free-grazing ducks and highly pathogenic avian influenza, Thailand. Emerg. Infect. Dis. 11:1664–1672.

Goloboff, P., S. Farris, and K. Nixon. 2005. TNT: Tree analysis using new technologies. www.zmuc.dk/public/phylogeny/TNT

Karesh, W., R. Cook, E. Bennett, and J. Newcomb. 2005. Wildlife trade and global disease emergence. Emerg. Infect. Dis. 11:1000–1002.

Kilpatrick, A. M., A. A. Chmura, D. W. Gibbons, R. C. Fleischer, P. P. Marra, and P. Daszak,2006. Predicting the global spread of H5N1 avian influenza. Proc. Natl. Acad. Sci. USA 103:19368–19373.

Kobasa, D., S. Kodihalli, L. Ming, M. Castrucci, I. Donatelli, Y. Suzuki, T. Suzuki, and Y. Kawaoka,1999. Amino acid residues contributing to the substrate specificity of the influenza a virus neuraminidase. J. Virol. 78:6743–6751.

Li, K., Y. Guan, J. Wang, G. Smith, K. Xu, L. Duan, A. Rahardjo, P. Puthavathana, C. Buranathai, T. Nguyen, A. Estoepangestie, A. Chaisingh, P. Auewarakul, H. Long, N. Hanh, R. Webby, L. Poon, H. Chen, K. Shortridge, K. Yuen, R. Webster, and J. Peiris. 2004. Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. Nature 430:209–213.

Lipatov, A., E. Govorkova, R. Webby, H. Ozaki, M. Peiris, Y. Guan, L. Poon, and R. Webster. 2004. Influenza: Emergence and control. J. Virol. 78:8951–8959.

Maddison, W. P. 1990. A method for testing the correlated evolution of two binary characters: Are gains or losses concentrated on certain branches of a phylogenetic tree? Evolution. 44:539–557.

Maddison, W. P., and Maddison, D. R. 2003. MacClade (version 4.06): Analysis of phylogeny and character evolution. Sinauer, Sunderland, Massachusetts.

Melville, D. S., and K. F. Shortridge. 2006. Spread of H5N1 avian influenza virus: An ecological conundrum. Lett. Appl. Microbiol. 42:435–437.

Normile, D. 2006a. Wild birds only partly to blame in spreading H5N1. Science 312:1451.

Normile, D. 2006b. South Korean flu mystery. Science 314:1371.

Obenauer, J., J. Denson, P. Mehta, X. Su, S. Mukatira, D. Finkelstein, X. Xu, J. Wang, Ma, J., Y. Fan, K. Rakestraw, R. G. Webster, E. Hoffmann, S. Krauss, J. Zheng, Z. Zhang, and C. Naeve. 2006. Large-scale sequence analysis of avian influenza isolates science. 311:1576–1580.

OIE (World Organization for Animal Health). 2006. Highly pathogenic avian influenza in the Republic of Korea. http://www.oie.int/eng/info/hebdo/ais_68.htm#Sec6

Olsen B., V. Munster, A. Wallensten, J. Waldenström, A. Osterhaus, and R. Fouchier, 2006. Global patterns of influenza a virus in wild birds. Science 21:384–388.

Rosenthal, E. 2006. Bird flu virus may be spread by smuggling. New York Times, April 15.

Salzberg, S., E. Ghedin, and D. Spiro. 2006. Shared data are key to beating threat from flu. Nature 440:605.

Shinya, K., S. Hamm, M. Hatta, H. Ito, T. Ito, and Y. Kawaoka. 2004. PB2 amino acid at position 627 affects replicative efficiency, but not cell tropism, of Hong Kong H5N1 influenza A viruses in mice. Virology 320:258–266

Shortridge, K. 1999. Poultry and the influenza H5N1 outbreak in Hong Kong, 1997: Abridged chronology and virus isolation. Vaccine 17:S26–S29.

Smith, G. J. D., X. H. Fan, J. Wang, K. S. Li , K. Qin, J. X. Zhang, D. Vijaykrishna, C. L. Cheung, K. Huang, J. M. Rayner, J. S. M. Peiris, H. Chen, R. G. Webster, and Y. Guan. 2006a. Emergence and predominance of an H5N1 influenza variant in China. Proc. Natl. Acad. Sci. USA 103:16936–16941.

Smith, G. J. D., T. S. P. Naipospos, T. D. Nguyen, M. D. de Jong, D. Vijaykrishna, T. B. Usman, S. S. Hassan, T. V. Nguyen, T. V. Dao,, N. A. Bui, Y. H. Leung, C. L. Cheung, J. M. Rayner, J. X. Zhang, L. J. Zhang, L. L. Poon, K. S. Li, V. C. Nguyen, T. T. Hien, J. Farrar, R. G. Webster, H. Chen, J. S. Peiris, and Y. Guan. 2006b. Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. Virology 350:258–268.

Stevens, J., O. Blixt, T. Tumpey, J. Taubenberger, J. Paulson, and I. Wilson. 2006. Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. Science 312:404–410.

Subbarao, E. K., W. London, and B. R. Murphy. 1993. A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. J. Virol. 67:1761–1764.

Van Borm, S., I. Thomas, G. Hanquet, B. Lambrecht, M. Boschmans, G. Dupont, M. Decaestecker, R. Snacken, and van den T. Berg,2005. Highly pathogenic H5N1 influenza virus in smuggled Thai eagles, Belgium. Emerg. Infect. Dis. 11**:**702–705.

Webster, R. G., W. Bean, O. Gorman, T. Chambers, and Y. Kawaoka.1992. Evolution and ecology of influenza A viruses. Microbiol. Rev. 56:152–179.

Webster, R. G., and Govorkova 2006. H5N1 influenza—Continuing evolution and spread. N. Engl. J. Med. 355:2174–2177.

Wheeler, W. C. 1996. Optimization alignment: The end of multiple sequence alignment in phylogenetics? Cladistics 12:1–9.

Wheeler, W. C., J. Delaet, D. Gladstein, and A. Varon. 2005. POY (version 3.012). Phylogeny reconstruction via optimization of DNA and other data; research.amnh.org/scicomp/projects/poy.php

WHO. 2006. Cumulative number of confirmed human cases of avian influenza A/(H5N1) reported to WHO. 1 March 2007. http://www.who.int/csr/disease/avian_influenza/country/cases_table_2007_03_01/en/index.html

First submitted 31 August 2006; reviews returned 23 November 2006; final acceptance 9 January 2007
Associate Editor: Rod Page