LA-UR- 09-06579

Approved for public release; distribution is unlimited.

Title:	Evolution of highly pathogenic avian H5N1 influenza viruses
Author(s):	Catherine A. Macken, Z#: 110192, T-6/T Division Margaret A. Green, Z#: 220880, T-6/T Division
Intended for:	Journal: PLOS Pathogens

Evolution of highly pathogenic avian H5N1 influenza viruses

Gabriele Neumann[1], Margaret A. Green[2], and Catherine A. Macken[2]

[1]Influenza Research Institute, Department of Pathobiological Sciences, University of Wisconsin-Madison, Madison, WI 53717; [2]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545;

Keywords: highly pathogenic avian H5N1 influenza viruses, evolution, two-time test, genotypes

Running title: Evolution of avian H5N1 influenza viruses

Abstract

Highly pathogenic avian H5N1 viruses have circulated in Southeast Asia for more than a decade, are now endemic in parts of this region, and have also spread to more than 60 countries on three continents. The evolution of these viruses is characterized by frequent reassortment events that have created a significant number of different genotypes, both transient and longer lasting. However, fundamental questions remain about the generation and perpetuation of this substantial genetic diversity. These gaps in understanding may, in part, be due to the difficulties of genotyping closely related viruses, and limitations in the size of the data sets used in analysis. Using our recently published novel genotyping procedure ('two-time test'), which is amenable to high throughput analysis and provides an increased level of resolution relative to previous analyses, we propose a detailed model for the evolution and diversification of avian H5N1 viruses. Our analysis suggests that (*i*) all current H5N1 genotypes are derived from a single, clearly defined sequence of initial reassortment events; (*ii*) reassortment of the polymerase and NP genes may have played an important role in avian H5N1 virus evolution; (*iii*) the current genotype Z viruses have diverged into three distinguishable sub-genotypes in the absence of reassortment; (*iv*) some potentially significant molecular changes appear to be correlated with particular genotypes (for example, reassortment of the internal genes is often paralleled by a change in the HA clade); and (*v*) as noted in earlier studies of avian influenza A virus evolution, novel segments are typically derived from different donors (i.e., there is no obvious pattern of gene linkage in reassortment). The model of avian H5N1 viral evolution by reassortment and mutation that emerges from our study provides a context within which significant amino acid changes may be revealed; it also may help in predicting the 'success' of newly emerging avian H5N1 viruses.

Introduction

Outbreaks of highly pathogenic avian influenza viruses of the H5N1 subtype were first reported in 1996 and 1997 in Southern China and Hong Kong (10,11,48,56). In recent years, these viruses have become endemic in poultry populations in parts of Asia, spread to Europe and Africa, and continue to transmit to humans with high mortality rates. The genetic features that account for the emergence, continued circulation, and spread of these viruses are currently not understood. Further, the high rate of genetic variation in these avian H5N1 viruses (6,51) makes the development of vaccines problematic, since the "target" is moving rapidly.

The genomic material of influenza A viruses is divided among eight segments, each coding for one or two proteins (reviewed in (41)). The hemagglutinin (HA) and neuraminidase (NA) proteins, each encoded by a single segment, define the antigenic properties of the virus. Currently, 16 HA (H1-16) and 9 NA subtypes (N1-9) are recognized (reviewed in (54)). The remaining six segments of influenza A viruses code for proteins with functions in viral replication, assembly and budding, and interference with host innate immunity (reviewed in (41)).

The segmentation of the influenza viral genome enables virus evolution by the process of *reassortment* when a single host cell is infected by at least two distinct viruses. Novel progeny viruses can be generated by swapping 1 - 7 segments of one virus for the corresponding segments of another virus. Reassortment produces evolutionary leaps, in contrast to the relatively continuous evolution by point mutation. Reassortment is critical for influenza virus evolution and accounts for the last three influenza virus pandemics, i.e., the 'Asian' pandemic in 1957, the 'Hong Kong' pandemic in 1968, and the current (2009) pandemic that is caused by swine-origin H1N1 viruses (18,39). The current pandemic strain originated from reassortment of Eurasian avian-like swine viruses with triple human/avian/swine H3N2 and H1N1 viruses that have circulated in pigs since 1997/1998 (18,39). Such dramatic examples of reassortment tend to overshadow the 'daily' contribution of reassortment to influenza virus

evolution, for example, among avian influenza viruses in their wild aquatic hosts (15), and among human influenza A viruses (38).

The highly pathogenic avian H5N1 influenza viruses that emerged in Southern China and Hong Kong in 1996/1997 possessed an A/goose/Guangdong/1/96-like HA gene (56) and shared internal segments with A/teal/Hong Kong/W312/97(H6N1)-like viruses (26) and A/quail/Hong Kong/G1/97(H9N2)-like viruses (23), suggesting that reassortment between H5N1, H6N1, and/or H9N2 viruses led to the emergence of highly pathogenic avian H5N1 viruses; however, recent data suggested that A/teal/Hong Kong/W312/97(H6N1)-like viruses may be descendants, rather than progenitors of highly pathogenic avian H5N1 influenza viruses (8). Although the outbreak in Hong Kong in 1996 was controlled by the depopulation of poultry in live poultry markets, the putative donor viruses continued to circulate (2,9,20,22,53). Prior to 2001, avian H5N1 viruses were isolated predominantly from geese (2,53). However, in early 2001, reassortants that had acquired some of the internal genes (i.e., the viral segments except HA and NA) from other avian viruses were isolated from terrestrial poultry in Hong Kong (19). Subsequent viral isolations indicated that H5N1 viruses reassorted multiple times with viruses of the H5N1, H7N1, H9N2, and H11N2 subtypes (4,5,14,19,21,29,44), until a predominant genotype (labeled genotype Z) was recognized in 2002 (21,29,44). Viruses of this genotype have, to date, persisted and spread geographically throughout Asia, Africa and Europe. The dominance of genotype Z is not absolute: non-genotype Z constellations have been observed but do not seem to persist and spread (14). Meanwhile, the HA and NA proteins of avian H5N1 viruses, while each derived from a single lineage, have evolved into distinguishable sub-lineages/clades (see http://www.who.int/csr/disease/avian_influenza/guidelines/nomenclature/en for a description of the clade nomenclature for HA).

A number of published studies on the reassortment of H5N1 viruses have identified and characterized a large number of distinct genotypes (4,5,14,19,21,29,44). However, significant gaps remain in our understanding of the generation and perpetuation of this substantial diversity. We, for example, do not fully understand the relationships among the distinct genotypes, and if their emergence and disappearance follows certain patterns, or is random. Moreover, the contribution of point mutations in the internal genes to the evolution of avian H5N1 viruses is not understood. These gaps in our understanding may be due to limitations in the size and/or composition of analyzed datasets, and/or difficulties in precisely defining a reassortant virus. To enable an accurate and detailed investigation of reassortment, we recently developed a novel statistical methodology, called the two-time test (33). Briefly, the two-time test acts like a sliding window over a time-period of interest. In each window of time, earlier viruses are regarded as potential donors to later 'test' viruses. By 'zooming in' on sliding time-windows, this methodology allows for a level of resolution and accuracy that has not been achieved with other approaches used for the genotyping of H5N1 viruses. Further, the structure of the two-time test allows for analysis of extensive data sets. This is critical since some genotypes may have a brief presence in viral populations and hence are less likely to be present in smaller datasets, leading to gaps in the understanding of evolution by reassortment.

Here, we implemented the two-time test for the evolutionary analysis of avian H5N1 viruses. The combination of our new methodology for genotyping with analysis of the now substantial dataset of complete H5N1 genomic sequences allowed a detailed and cohesive description of the contributions of reassortment and point mutation to the emergence of the current H5N1 viruses. The most important findings are: (*i*) the nature and sequential order of reassortment events that contributed to current H5N1 viruses, as well as the lineages of donor viruses that contributed gene segments to H5N1 virus evolution, can be established; (*ii*) despite the generation of multiple genotypes due to frequent reassortment during the early phase of

H5N1 circulation, all current genotypes are derived from a single sequence of initial reassortment events; (*iii*) reassortment of the polymerase and NP segments may have been critical for avian H5N1 virus evolution; (*iv*) current H5N1 viruses are evolving primarily by point mutations, rather than reassortment; (*v*) the accumulation of point mutations in recent H5N1 virus has led to the formation of distinguishable sub-genotypes of genotype Z; and (*vi*) some molecular changes of potential significance (such as changes in the HA clade) are correlated with particular genotypes. Collectively, these findings may help in identifying amino acid changes that affect viral protein functions. The identification of amino acid changes that affect protein functions is a challenge in the field. As we suggest here, variation among influenza viral isolates that appears random when viewed in aggregate (such as in sequence alignments) may contain significant patterns when viewed in the context of a detailed model of evolution by reassortment and mutation. Thus, our analysis leads to testable hypotheses about amino acid changes that may have affected avian H5N1 virus evolution.

Materials and Methods

(i) Data

The study was limited to avian influenza A viruses with > 90% of full-length sequences for each segment in the genome (henceforth described as a whole genome). The study genotyped highly pathogenic avian influenza A (H5N1) viruses from 2000 to 2008 ('test viruses'). Detailed genotyping by the two-time test of H5N1 viruses from 1996 - 1999 was not possible because of the paucity of genetic data from closely related, earlier viruses to act as reference viruses. Avian influenza viruses of all subtypes (including H5N1) from Asia, Africa and Europe were regarded as potential donors for reassortment events; these viruses therefore served as 'reference viruses'. Genetic data were drawn from Genbank (1) with the most recent update of the analyses occurring in May 2009.

(ii) Alignments and phylogenetic trees

Alignments of nucleotide sequences were generated using the profile alignment option of CLUSTALW, with manual adjustment if necessary. Profiles were hand-curated alignments of representative sequences, which included all insertion and deletion patterns present in avian influenza A viral sequences in the database. The two-time test uses the HKY model of evolution in neighbor-joining trees, in order to achieve high-throughput (in our earlier study (33), we demonstrated that results of the two-time test were robust to the method of tree inference). Some additional analyses used the GTR + G + I model of evolution (16); these were fitted by maximizing likelihood using PhyML (24).

(iii) Two-time test

The two-time test for genotyping influenza viruses was described in detail in our earlier publication (33). 'Test' viruses in one time period are genotyped relative to 'reference' viruses in an earlier, neighboring time period. 'Reference' phylogenetic trees - one for each segment - of the viruses in the earlier time period are constructed so that only one virus of each distinct genotype, having high bootstrap support (>90%) for placement in the phylogeny, is included. Test viruses are added, each by itself, to the reference trees and bootstrap values recalculated for the augmented tree. If a segment of a test virus has bootstrap support >70% for being the sister to one of the reference viruses, then the lineage of this segment of the test virus is classified as being that of the reference virus. By repeating this process for all segments, the genotype of a test virus is classified in terms of the reference viruses, which represent donor lineages. We then infer the most parsimonious exchange of segments that relates genotypes of test viruses to each other and to the genotypes of reference viruses. The resultant "evolutionary paths" model the evolution by reassortment of the test viruses.

(iv) Genotyping

We genotyped the PB2, PB1, PA, NP, M and NS segments of all H5N1 viruses in our dataset from 2000 to 2008 (as stated earlier, genotyping of H5N1 viruses from 1996 - 1999 was not possible because of the paucity of genetic data from closely related, earlier viruses of H5N1 and other serotypes to act as reference viruses). To study the relationship between the genotype of internal segments and the HA gene, which has a potentially significant effect on viral evolution due to the immune response, we identified the clade of the HA segment of each H5N1 virus according to the WHO nomenclature (http://who.int/csr/disease/avian_influenza/H5CompleteTree.pdf). Because some of our 'test' viruses are not included in this tree, we inferred the HA phylogeny of all test viruses under the HKY model of evolution with maximum likelihood optimization using PhyML (24). If we found strong support for the association of a virus of known clade with a virus with an unclassified HA, the respective clade number was assigned. Otherwise, the HA was not classified.

Results

For a comprehensive analysis of avian H5N1 virus evolution, we genotyped avian H5N1 viruses for which complete, or almost complete, genomic sequences are available. Human isolates were not included to rule out the distortion of data sets with mutations that may reflect adaptation to mammalian species. In total, 634 H5N1 viruses were genotyped and analyzed for their evolutionary relationships. Potential donor viruses included avian influenza A viruses of unrestricted subtypes isolated since 1979 in Asia, Africa, and Europe, i.e., geographical regions that have experienced H5N1 virus outbreaks. Genotyping was carried out using the two-time test (33). This test was originally developed for the genotyping of the components of the replication complex (i.e., the PB2, PB1, and PA genes that encode the respective subunits of the polymerase complex, and the nucleoprotein gene) of avian influenza A viruses of unrestricted subtype and unrestricted geographical region. Modifications needed to apply the two-time test to the closely related avian H5N1 viruses are explained in the supplementary

material [S.I]. Although the two-time test for genotyping is more difficult to implement than genotyping using BLAST (as in, for example (32)), the results are more precise; this extra precision is important for inferring a detailed model of evolution.

The two-time test acts like a sliding window in which earlier viruses become potential donors to later 'test' viruses. With each successive time window, the reference viruses are updated, thus ensuring that they reflect significant recent evolutionary events. Because of the frequent reassortment and rapid rate of variation in the avian H5N1 viruses, if one runs the two-time test using reference viruses that do not closely represent recent evolution, the test will lose power and bootstrap support for classifying segments to lineages will deteriorate. In our current study of avian H5N1 virus evolution, we needed to analyze the dataset of viruses from 2000 - 2008 in three time periods, in order to maintain reference sets that captured significant phases of evolution. Our first analysis period included avian H5N1 viruses from 2000 - 2001; frequent reassortment has been reported for this period (5,14,19,20,29). (Note that H5N1 viruses from 1996 - 1999 could not be genotyped by the two-time time test because of the paucity of genetic data from closely related, earlier viruses of H5N1 and other serotypes to act as reference viruses.) Our second period included viruses from 2002 - 2004, with reference viruses from prior to 2002; this is the period during which the dominance of genotype Z in Southeast Asia was consolidated. Our third period, 2005 - 2008, with reference viruses from prior to 2005, has been one of extensive geographical spread of avian H5N1 viruses.

Analysis of avian H5N1 virus evolution from 2000 - 2001

For the first period, candidate 'reference' viruses were all available Asian, European, and African avian influenza viruses isolated between 1979 and 1999, including avian H5N1 viruses from 1996-1999.

The set of 'test' viruses for the 2000 - 2001 time period is sparse with complete genomic sequences for only 36 avian H5N1 viruses. It was possible to unequivocally assign almost all

segments of each test virus to a lineage represented by a reference virus, thereby establishing the genotype of the test virus. Bootstrap support for association with a reference lineage was almost always at least 90%. Four PB2 segments had bootstrap support below 90%; three of those had bootstrap support above 80%. Two NP and two M segments could not be assigned a lineage, presumably because the donor of the reassorted segments was not present in the database and hence could not be included in the reference set. The small dataset of candidate reference viruses means that the genotyping of 2000 - 2001 H5N1 viruses lacks the resolution of later time periods. While we can distinguish genotypes with high confidence, a larger reference set would allow a closer pinpointing of the donor viruses, and finer resolution classification of viruses into genotypes.

In our analyses, two avian H5N1 viruses (A/goose/Guangdong/1/96 and A/duck/Guangxi/07/99; genotypes 1 and 2, respectively, in Fig. 1) represent the starting points for subsequent H5N1 reassortment events. These viruses differ significantly only in their NS gene allele: A/goose/Guangdong/1/96 possesses an allele B NS segment (56), while the 1997 Hong Kong H5N1 viruses and almost all contemporary H5N1 viruses contain an NS segment of allele A (20,56). The common progenitor virus (shown in dotted lines in Fig. 1) is not known, although the chronology of events suggest that the common ancestor is more likely to have an allele B rather than an allele A NS gene.

As described by others (5,14,19,20,29), the 2000 - 2001 time period was characterized by frequent reassortment events (Fig. 1); the novel genotypes involved genetic lineages represented by eight different reference viruses from four subtypes (Fig. 1). These novel genotypes were typically represented by a very small number of viruses (see the box in Fig. 1 that lists the number of viruses per genotype), and were replaced rapidly. Our detailed analysis also allowed us to establish evolutionary relationships among the genotypes. The most parsimonious model is shown in Fig. 1; notably, reassortment typically involves replacement of a single segment at a time (for a summary of avian H5N1 virus evolution, see also Fig. 4). Four

genotypes represented in Fig. 1 differed by more than one segment from all other genotypes. This finding suggests that intermediate viruses are missing, or that multiple segments derived from multiple viruses were exchanged during a single reassortment event. As observed in our earlier analysis of the replication complex (PB2, PB1, PA and NP) of avian influenza A viruses (33), reassorting segments were usually derived from different donor lineages. This behavior has also been noted implicitly or explicitly in other studies (5,14,15,29).

Our maximum parsimony inference leads to a model for the series of events leading from A/goose/Guangdong/1/96 and A/duck/Guangxi/07/99 (genotype 1 and 2, Fig. 1) to genotype Z viruses (i.e., genotype 14 in our analysis). From the model of Fig. 1, we identified A/duck/Guangxi/50/01 (the representative virus for genotype 14 in Fig. 1) and A/chicken/Hong Kong/NT873.3/01 (also a member of genotype 14, but not listed by name in Fig. 1), as the two earliest known genotype Z viruses. This is consistent with findings by Duan *et al.* (14) who also identified A/duck/Guangxi/50/01 as the earliest known genotype Z virus. Interestingly, A/chicken/Hong Kong/NT873.3/01 is the last virus in our inferred evolutionary path leading to contemporary H5N1 viruses that does not have a deletion in the NA gene. A/duck/Guangxi/50/2001, along with all other genotype Z viruses and their descendents has a 20 amino acid deletion in the stalk region of the NA (amino acids 49-68 (19)). Further, the HA of A/chicken/Hong Kong/NT873.3/01 is from clade 3, while the HA of A/duck/Guangxi/50/01 is from the genetically distinct clade 9. Currently, it is not known if the substitution of the HA clade and/or the deletion in NA were critical for the evolution of avian H5N1 influenza viruses.

Noticeable in early H5N1 virus evolution are frequent reassortment events of the PB2, PB1, PA, and NP genes (which form the viral replication machinery). For example, replacement of the NP gene of genotype 12 viruses resulted in four different genotypes (i.e., genotypes 10, 11, 13, and 14, Fig. 1). This 'outburst' in NP reassortment may suggest that the NP segment of genotype 12 viruses was not optimal in the genetic background of this genotype. Three of the resulting genotypes did not persist; however, replacement of the genotype 12 NP segment with

an NP segment most closely related to the NP segment of an H9N2 virus from 1998 founded genotype Z and hence all currently circulating H5N1 viruses. Another noticeable event in early H5N1 virus evolution is a 5-amino acid deletion in the NS1 protein that is characteristic of all recent avian H5N1 viruses (19). In our maximum parsimony model of Fig. 1, the same 5-amino acid deletion occurs twice. One equally parsimonious model exists: it would swap the positions of genotypes 7 and 8, leading to a single NS1 deletion event but repeated reassortments of PB2. While both repeated events are possible, more likely this repetition is a result of the sparse early data set. The uncertainty in these early events does not affect the fundamental insights of our model.

As a side lineage in avian H5N1 virus evolution, replacement of the PB2, PB1, and PA segments of A/duck/Guangxi/07/99 virus (genotype 2) led to genotype 3, represented by A/duck/Anyang/AVL-1/01 (Fig. 1). From our analysis, it is not clear if several viral segments reassorted in parallel, or if intermediate viruses are missing due to the sparse reference set for this time period. Genotype 3 viruses have been described by others as genotype X0 and have been noted to originate from reassortment events separate from those that founded genotype Z (14). These viruses underwent several reassortment events in the period 2001 to 2002 (genotypes 4-6 in Fig. 1, and genotype 25 - 30 in Fig. 2); genotypes 6, 25, 26, 29 and 30 have been described as the transient genotypes X5-7 and X9-10 (14). A/duck/Anyang/AVL-1/01-like viruses were present in 2001 - 2004, but have not been detected since.

Most early H5N1 influenza viruses possessed an HA protein of clade 0 (see box in Fig. 1 that lists the HA clades with a representative virus of each genotype; representative viruses are randomly chosen from the first year that a given genotype was recognized). However, as stated earlier, the two earliest known genotype Z viruses possessed HA proteins from the genetically distinct clades 3 and 9.

Analysis of avian H5N1 virus evolution from 2002 - 2004

Our genotyping and maximum parsimony inference leads to a model of evolution for this time period described in Fig. 2. (Note that an equally parsimonious model has genotype 18 following from genotype 21.) In this model, only two H5N1 genotypes from the previous period contributed to the creation of novel genotypes in the 2002 - 2004 time period, namely genotype 14, represented by A/duck/Guangxi/50/01 (Figs. 1 and 2), and the A/duck/Anyang/AVL-1/01 genotype (genotype 3 in Figs. 1 and 2). As stated earlier, A/duck/Anyang/AVL-1/01-like viruses underwent several reassortment events. None of the resultant novel genotypes was represented by more than one virus or founded new lineages: They constitute some of the transient 'X-series' described by Duan *et al.* (14). The A/duck/Anyang/AVL-1/01 genotype was last detected in 2004.

The majority of viruses from this test period (120 out of 183) were classified as genotype Z viruses, attesting to the dominance of this genotype during the 2002 - 2004 time period. Compared to the earliest known genotype Z viruses, genotype Z viruses have now evolved by point mutations; this evolution is depicted by a different line style (compare genotype 14 in Figs. 1 and 2). While reassortment of genotype Z viruses occurred, the resulting novel genotypes were mostly represented by a small number of viruses and were not detected after 2004. Similarly to the 2000 - 2001 time period, most reassortment events involved the polymerase and NP genes. We, however, observed an interesting change in the reassortment pattern between the two consecutive test periods: during the 2000 - 2001 time period, most donor viruses were of sub-types other than H5N1; by contrast, during the 2002 - 2004 time period, most donor viruses were of the H5N1 subtype. Note that occasional reassortment with H6N1 viruses (see Fig. 2), and other unclassifiable and presumably novel viruses (not shown) was detected, suggesting that the opportunity for reassortment with non-H5N1 viruses continued to exist. This change in the reassortment pattern may suggest a relatively high biological 'fitness' of genotype Z viruses, so that the introduction of more distantly related genes from non-H5N1 viruses may not have provided a selective advantage.

By 2004, at least 8 different HA clades/subclades (1, 2.1.1, 2.3.1, 2.3.2, 2.4, 6, 8 and 9; with clade 1 being most frequent) were detected on the backbone of the genotype Z internal genes. This variety of HA clades in association with a single genotype (Z) is in contrast to the limited diversity of HA clades in conjunction with most other genotypes in our analysis, suggesting a burst of evolution and diversification of HA on the novel constellation of internal genes.

Nineteen viruses are not represented in Fig. 2, for the sake of clarity of the figure. Six of these were reassortants of genotype Z that involved at least one segment derived from a novel donor lineage, that is, a lineage that was not represented in the database at the time of the two-time testing. (Subsequent BLAST searches of the database also failed to reveal any probable donors.) These viruses provide evidence of the continued introduction of novel genetic material into the H5N1 population, an observation also made by others (3,14,29,51). Four viruses collected in 2004 from tree sparrows in Henan, China, differed greatly from our reference viruses in most segments and hence could not be placed in an evolutionary path. The remaining viruses had segments for which bootstrap values in the two-time test were below the acceptable threshold, or were themselves of uncertain origin, since they had either an extremely close match to 1997 viruses or an unusual pattern of amino acid replacements. None of these 19 viruses had descendents in our data set.

Analysis of avian H5N1 virus evolution from 2005 - 2008

This period includes the largest number of virus genomes (415) for analysis. Our genotyping and maximum parsimony inference leads to a model of evolution for this time period described in Fig. 3. The most important finding is the divergence of genotype Z viruses into three distinguishable sub-genotypes, referred to as Z.1 - Z.3 (Fig. 3; see S.I for details of ascertaining sub-genotypes). Sub-genotype Z.1 (with the representative virus A/peregrine falcon/Hong Kong/D0028/04) includes the Qinghai-Lake viruses and their descendents, which

have spread into Europe and Africa. Sub-genotype Z.2 (with the representative virus A/chicken/Yunnan/1252/03) is composed of avian H5N1 viruses isolated in Thailand and Vietnam. Sub-genotype Z.3 (with the representative virus A/duck/Hunan/782/03) is composed of 13 avian H5N1 viruses, most of which were isolated from chickens in Indonesia.

The three sub-genotypes differ from each other and from genotype Z by phylogenetically distinguishable PB2, PB1, and PA genes; some mutations that distinguish these genes result in amino acid replacements (in Fig. 3, the diversification into sub-genotypes is represented by different line styles of the same color). In contrast, no such diversification exists for the NP and M segments. The NS segment of sub-genotype Z.1 viruses could be distinguished phylogenetically from the NS segments of sub-genotype Z.2 and Z.3 viruses, which could not be distinguished from each other. The divergence into three sub-genotypes is paralleled by differences in the HA clade. All except one genotype Z virus (genotype 14) possess an HA gene of clade 1, which is maintained in sub-genotype Z.2 viruses; sub-genotype Z.1 viruses possess an HA gene of clade 2.2(.1), with seven exceptions; and sub-genotype Z.3 viruses have HA genes of clades 2.1.2 and 2.1.3.

Sub-genotype Z.1 and Z.2 viruses persisted and expanded, but did not undergo reassortment during 2005 - 2008. By contrast, sub-genotype Z.3 viruses were rapidly replaced by reassortment events (Fig. 3). As observed during the early phase of avian H5N1 virus evolution, most reassortment events involved components of the replication machinery. Some of these reassortment events coincided with potentially significant molecular changes. For example, replacement of the PA and NS segments of sub-genotype Z.3 viruses led to genotype 31 viruses, which are characterized by a C-terminally truncated NS1 protein (lacking the PDZ domain motif (40)) and a mutation in PB1 that results in a truncated PB1-F2 protein, which has recently been identified as a determinant of virulence (12). Descendents of genotype 31 viruses, i.e., genotype 34 viruses, encode a full-length PB1-F2 protein but a C-terminally deleted NS1 protein. To date, the biological significance of these findings is unknown. Reassortment

generating genotypes 31 and 34 is again mirrored by changes in the HA clade from clade 2.1 (sub-genotype Z.3) to clade 2.3 in genotypes 31 and 34.

Clade 2.3.4 (Fujian-like) viruses were originally described as genotype Z (44) but later reclassified as genotype V (14). Recently, clade 2.3.4 viruses were also observed in genotype G (46). Our analysis provides a more detailed picture of the evolution of clade 2.3.4 viruses, which form the majority of genotypes 31, 32, 33, 35, and 36 (Fig. 3). Genotype V is represented by our genotype 33 and descendants (35 - 37). According to earlier studies, genotypes V and Z differ in the PA segment (5), a reassortment event that we also detect (see sub-genotype Z.3 to genotype 33, Fig. 3). In addition, we detect a reassortment event in the PB2 gene, resulting in genotype 35 which has persisted and expanded, and a transient genotype 36. Genotype G viruses are represented in our analysis by genotypes 32 and 34. Duan *et al.* reported that genotypes Z and G differ in the PB2 segment (14). While we detect reassortment events in PB2 leading to genotypes 32 and 34, we also detect additional reassortment of PA and NS in genotype 34 viruses as compared to genotype 32 viruses.

Interestingly, sub-genotype Z.1 and genotype 35 viruses (which possess closely related PB2 genes) have circulated and geographically expanded since 2005. Currently, it is not known if the PB2 gene possessed by viruses of sub-genotype Z.1 and genotype 35 contributed to the 'success' of these groups of avian H5N1 viruses.

Sixteen viruses are not shown on the evolutionary paths of Fig. 3. Seven of these contained between 1 and 3 apparently novel segments (i.e., segments that differed substantially from all sequences in the database at the time of analysis, which is presumed to be evidence of the continued introduction of novel genetic material into the H5N1 avian virus population); five contained segments that could not be classified because of low bootstrap support in our two-time test; three were very similar to H5N1 viruses from 1997 and hence their origin is uncertain; and one virus from a wild duck differed from all viruses on the evolutionary paths by more than

three segments, arising from possibly multiple, simultaneous reassortment events or unobserved intermediates.

Discussion

Here, we present a detailed and comprehensive analysis of the evolution of highly pathogenic avian H5N1 influenza viruses. Our genotyping analysis allows us to propose a model that summarizes the emergence of the currently dominant groups of viruses, i.e., sub-genotype Z.1 viruses (HA clade 2.2), sub-genotype Z.2 viruses (HA clade 1), and genotype 35 viruses (HA clade 2.3.4). Our major findings are: (*i*) all current avian H5N1 viruses appear to be derived from one clearly defined sequence of reassortment events; (*ii*) the accumulation of point mutations in recent avian H5N1 viruses has led to the formation of distinguishable sub-genotypes of genotype Z; (*iii*) reassortment of an internal gene segment is often paralleled by a change in the HA clade; (*iv*) the eight influenza viral RNA segments do not contribute equally to reassortment, in contrast to findings in earlier studies of avian influenza A viruses of mixed subtypes (15,33). Overall, our model provides a more detailed picture than previously available of the evolution of avian H5N1 virus and points to a mixture of critical reassortment events and significant evolution by point mutation in the generation of the current diversity of this viral population (see Fig. 4).

A number of studies have addressed the evolution of H5N1 influenza viruses (4,5,14,19,21,29,44). These studies provided phylogenetic analyses of the individual viral segments and tentatively described the evolutionary relationships of some of the major genotypes. Consistent with these reports, we find that a switch in the NS allele occurred early in H5N1 virus evolution. A/goose/Guangdong/1/96 virus, which has been described as a progenitor of current H5N1 viruses (56), is characterized by an allele B NS segment, while almost all of the current H5N1 viruses possess NS segments of allele A. Here, we identify A/duck/Guangxi/07/99 virus as the earliest avian H5N1 virus with the internal genes of

A/goose/Guangdong/1/96 but an allele A NS segment. Hence, we propose A/duck/Guangxi/07/99 virus as the prototype H5N1 virus from which all current H5N1 viruses descended.

Duan *et al.* (14) speculated that "genotype B virus was probably the progenitor of genotype Z viruses". This is consistent with our analysis since we find that genotype B viruses are members of genotypes 9 and 12. Our detailed analysis, however, recognizes a reassortment event in the PA gene that separates genotype B viruses into (at least) two different genotypes.

After their appearance in 2001, genotype Z viruses underwent frequent reassortment events, predominantly with H5N1 viruses. A few reassortment events involved non-H5N1 viruses, which is congruent with observations that avian influenza viruses of other subtypes likely co-circulated with H5N1 viruses in Southeast Asia at that time (8,55). Notably, these reassortment events produced only transient genotypes. The observation that these novel genotypes did not replace genotype Z suggests that genotype Z was highly competitive and that the introduction of novel genes from H5N1 or non-H5N1 viruses did not provide a selective advantage.

The persistence of genotype Z viruses resulted in the accumulation of point mutations that led to three distinguishable sub-genotypes (designated here as Z.1, Z.2, and Z.3). These sub-genotypes have not been recognized in other studies, although H5N1 viruses isolated from Vietnam and Thailand have been described as a geographically distinct group within genotype Z viruses (45). The divergence into sub-genotypes is apparent for the polymerase and NS genes, but not for the NP and M genes. This finding suggests that the accumulated point mutations are not merely the result of the error rate of the viral replication complex, but, at least in part, reflect evolutionary pressure.

The polymerase proteins of sub-genotype Z.1, Z.2, and Z.3 viruses differ by relatively small numbers of amino acid changes, some of which are notable. For example, the consensus

sequence of sub-genotype Z.1 virus PB2 proteins differs by nine amino acids from that of sub-genotype Z.3 virus PB2 proteins. At six of these positions, sub-genotype Z.1 PB2 proteins possess amino acids that are typically not found in avian influenza viruses (among those is Lys at position 627 which is now recognized as a determined of pathogenicity in mammalian species (25,47)). Three of these rare changes (not including Lys at position 627) were also found in the PB2 proteins of genotype 35 viruses, reflecting the introduction by reassortment of a sub-genotype Z.1-like PB2 segment into genotype 35 viruses (Fig. 3). Since both groups of viruses have circulated for several years, one may speculate that the unique amino acids in their PB2 proteins are of a biological significance by, for example, increasing the replicative ability. However, experimental data that test this hypothesis are not available at this point.

The HA proteins of avian H5N1 viruses have diverged into multiple clades and sub-clades (see http://www.who.int/csr/disease/avian_influenza/guidelines/nomenclature/en). Interestingly, our analysis suggests that changes in the HA clade often coincide with changes in the make-up of the internal genes. (It is important to recall that HA was not used in determining genotypes.) Examples of such coincidence occur throughout our modeled evolutionary paths, and include: the transition from genotype 14 (clade 1 HA) to genotype 16 (clade 2.5 HA) (Fig. 1), the transition from genotype 17 (clade 2.4 HA) to genotype 18 (clade 5 HA) (Fig. 1), or the earlier mentioned examples that involve the separation of genotype Z into sub-genotype Z.1 - Z.3 and the transition of subtype Z.3 to genotype 31 and from there to genotype 34 (Fig. 3). One might speculate that the HA proteins of different clades/sub-clades differ in their affinity to cellular receptors, and that lower receptor-binding affinity could be compensated for by an efficient replication complex; similarly, a replication complex with low efficiency could be compensated for by an HA protein with high receptor-binding affinity.

Our model of avian H5N1 influenza virus evolution points to an important role of the polymerase and NP genes for the evolution of these viruses, in contrast to earlier studies of mixed subtypes that suggested that all segments may participate equally in reassortment

(15,33). Our genotyping showed that most reassortment events involved components of the polymerase complex. Moreover, the polymerase proteins of genotype Z viruses continued to evolve through point mutations, resulting in sub-genotypes Z.1 - Z.3. The polymerase complex is now recognized as a determinant of pathogenicity (17,25,30,42,50,52), although neither the contributions of the individual polymerase proteins nor the underlying mechanisms are fully understood. Overall, the currently available data suggest that differences in the replicative ability may account for the observed differences in virulence (13,34,37,43). Replacement of the polymerase genes in the early phase of avian H5N1 virus evolution, or point mutations in these genes in recent avian H5N1 viruses may thus have affected the replication levels of the respective viruses and hence their ability to compete with co-circulating viruses of other genotypes or subtypes.

The NS gene reassorted less frequently than the polymerase and NP genes; however, it was involved in several notable events in avian H5N1 virus evolution. Early in avian H5N1 virus evolution, a switch from allele B to allele A occurred (Fig. 1). In avian species, NS genes of both alleles are found (28). In contrast, human influenza viruses carry exclusively allele A NS genes (28). Thus, although unproven, the acquisition of an allele A NS segment may have contributed to the ability of avian H5N1 viruses to infect humans. A second noticeable event involving the NS gene was the emergence of a 15-nt deletion that resulted in the deletion of amino acids 80-84 of NS1. This deletion is found in all contemporary avian H5N1 viruses (Fig. 1) and a recent study suggested that it increases virulence (31). This deletion may have occurred in conjunction with the initial reassortment of the PB2 gene segment on the path to emergence of genotype Z (Fig. 1). A third noticeable event occurred in recent avian H5N1 virus evolution when a mutation in the genotype 31 NS gene created a premature stop that abrogated expression of the two C-terminal amino acids of NS1. The four C-terminal amino acids of NS1 form the PDZ ligand domain motif (40) which contributes to virulence (27). An influenza A/WSN/33 virus lacking the 4 C-terminal amino acids of NS1 was attenuated *in vitro* and *in vivo* and did not kill mice (27).

However, a number of avian H5N1 viruses have now been isolated that lack an intact PDZ domain [all genotype 31 viruses (n=7) and all genotype 34 viruses (n=38)]. These viruses were isolated from chickens and ducks in Southern China, where avian H5N1 viruses of other genotypes (which encoded full-length NS1 proteins) circulated at the same time. These findings suggest that avian H5N1 influenza viruses lacking a fully functional PDZ domain are viable in nature.

The truncation of the PDZ domain in genotype 31 viruses was accompanied by a mutation in the PB1 gene that created a premature stop codon in the PB1-F2 protein, resulting in a truncated version of 25 amino acids. PB1-F2 functions as a pro-apoptotic factor (7,57), enhances inflammation in mice, and increases the severity of secondary bacterial infections (36). Moreover, it interacts with the PB1 protein to retain it in the nucleus and may through this mechanism affect virulence (35). Currently, it is not clear if the C-terminal truncation of NS1 and the truncation of PB1-F2 are functionally linked. However, genotype 34 viruses possess a truncated NS1 protein while encoding a full-length PB1-F2 protein. Hence, abrogation of PB1-F2 expression is not required for the viability of viruses lacking a fully functional PDZ domain.

The identification of amino acid changes that affect protein functions remains a challenge in the field. Random mutagenesis approaches are cumbersome and their outcome is unpredictable. In the past, researchers often resorted to the comparison of pairs of viruses that differ in their virulence or pathogenicity. These approaches have identified amino acid changes that affect virulence and/or pathogenicity in the genetic background of the test viruses; however, these changes may not have similar effects in the genetic context of other viruses (for example, discussed in (49)); the glutamic acid-to-lysine replacement at position 627 of many highly pathogenic avian H5N1 viruses is a notable exception and is now recognized as a general determinant of virus pathogenicity in mammalian species. Recently, large-scale sequencing approaches have been employed to identify signature amino acid changes in viral genomes. Such an approach identified the PDZ domain in the NS1 protein (40); however, sequence

alignments do not take into account the genetic relationships among viruses and hence fail to identify changes that correlate with reassortment events or diversification through point mutations. By contrast, our detailed genotyping approach identifies clearly defined groups of viruses that differ by small numbers of amino acid changes. These changes can now be tested for their biological significance; examples include the changes in HA clade, the amino acid difference among sub-genotype Z.1 – Z.3 polymerase proteins, the amino acid differences between Z.1 and Z.2/Z.3 NS1 proteins, the significance of the truncated PB1-F2 and NS1 proteins (genotype 31), and the significance of the co-occurrence of these events.

In summary, our studies lead us to a rich model of avian H5N1 virus evolution, involving reassortment and significant gene diversification by point mutations. Such knowledge may help in predicting future events in avian H5N1 virus evolution; for example, reassortment of the PB2 gene appears to be much more likely to found a long-lasting genotype than reassortment of the M gene. In addition, our detailed model of avian H5N1 virus evolution identifies amino acid changes that may be critical for the function of the viral protein; these changes may have escaped earlier detection due to a lower level of resolution in evolutionary studies, or because analyses of sequence alignments were not correlated with phylogenies. Finally, we wish to emphasize that the two-time test is also suitable for the analysis of other groups of influenza viruses, such as pandemic H1N1 viruses, or seasonal H1N1 or H3N2 viruses.

Figure Legends

Figure 1. Model of the evolution of avian H5N1 influenza viruses from 2000 – 2001. Depicted are the 'internal' genes in the order PB2, PB1, PA, NP, M, and NS (six bars, from top to bottom). The color of a bar represents the lineage from which a segment was inferred to have originated. For example, lime green NS segments are inferred to have originated from the lineage represented by A/duck/Guangxi/07/99. Color changes indicate reassortment events. 'ΔNS' and a triangle in the NS segment depict an internal deletion in the NS segment. Dotted arrows represent uncertainty in the NS allele of the most recent ancestor of the contemporary lineage of H5N1 viruses. 'Dotted' viruses are not in the databases, but can be inferred. Each genotype was assigned a number. For each genotype, a representative virus, the number of viruses per genotype, and the dominant HA clade are listed. Also listed is the history of virus generation; for example, genotype 12 evolved from genotype 9 by a reassortment event that resulted in the introduction of an A/duck/Hong Kong/Y439/97(H9N2)-like PA segment (shown in light blue).

Figure 2. Model of the evolution of avian H5N1 influenza viruses from 2002 – 2004. The details of bars, colors, NS deletion and descriptions of genotypes are as described in Fig. 1. In this period, evolution by point mutation is indicated by changes in line style. For example, compare genotype 14 in Figs. 1 and 2. Different line styles in the same color for a given segment in Fig. 2 indicate divergence from the original representative of the lineage in Fig. 1. In another example, the NS genes of genotypes 14 and 21 differ significantly. Both are lime green (describing a common ancestor, A/duck/Guangxi/07/1999). However, the NS gene of genotype 21 is closely related to A/duck/Zhejiang/52/00, while that of genotype 14 is closely related to A/duck/Guangxi/50/01. The patterned wedge labeled with "120 viruses" indicates the persistence and expansion in numbers of genotype 14 viruses. NC; not classified.

Figure 3. Model of the evolution of avian H5N1 influenza viruses from 2005 – 2008. The details of bars, line styles, colors, NS deletion, and descriptions of genotypes are as described in Figs. 1 and 2. Genotype Z viruses have now differentiated into three distinguishable sub-genotypes, depicted by different line styles. The triangle in PB1 indicates a truncated PB1-F2 protein, while the second triangle in NS indicates a C-terminally deleted NS1 protein (see text for more details). The patterned wedges indicate the persistence and expansion in numbers and/or geographic origin of the respective (sub-)genotypes.

Figure 4. Summary of the model of evolution of avian H5N1 influenza viruses from 2000-2008. Shown is a summary of the evolutionary events that led to the currently dominant HA clade 1, 2.2, and 2.3.4 viruses.

References

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL(2008) GenBank. *Nucleic Acids Res* **36**: D25-D30.

2. Cauthen AN, Swayne DE, Schultz-Cherry S, Perdue ML, Suarez DL(2000) Continued circulation in China of highly pathogenic avian influenza viruses encoding the hemagglutinin gene associated with the 1997 H5N1 outbreak in poultry and humans. *J Virol* **74**: 6592-9.

3. Chen H, Deng G, Li Z *et al.*(2004) The evolution of H5N1 influenza viruses in ducks in southern China. *Proc Natl Acad Sci U S A* **101**: 10452-7.

4. Chen H, Li Y, Li Z *et al.*(2006) Properties and dissemination of H5N1 viruses isolated during an influenza outbreak in migratory waterfowl in western China. *J Virol* **80**: 5976-83.

5. Chen H, Smith GJ, Li KS *et al.*(2006) Establishment of multiple sublineages of H5N1 influenza virus in Asia: Implications for pandemic control. *Proc Natl Acad Sci U S A* **103**: 2845-50.

6. Chen R, Holmes EC(2006) Avian influenza virus exhibits rapid evolutionary dynamics. *Mol Biol Evol* **23**: 2336-41.

7. Chen W, Calvo PA, Malide D *et al.*(2001) A novel influenza A virus mitochondrial protein that induces cell death. *Nat Med* **7**: 1306-12.

8. Cheung CL, Vijaykrishna D, Smith GJ *et al.*(2007) Establishment of influenza A virus (H6N1) in minor poultry species in southern China. *J Virol* **81**: 10402-12.

9. Chin PS, Hoffmann E, Webby R *et al.*(2002) Molecular evolution of H6 influenza viruses from poultry in Southeastern China: prevalence of H6N1 influenza viruses possessing seven A/Hong Kong/156/97 (H5N1)-like genes in poultry. *J Virol* **76**: 507-16.

10. Claas EC, de Jong JC, Van Beek R, Rimmelzwaan GF, Osterhaus AD(1998) Human influenza virus A/HongKong/156/97 (H5N1) infection. *Vaccine* **16**: 977-8.

11. Claas EC, Osterhaus AD, Van Beek R *et al.*(1998) Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *Lancet* **351**: 472-7.

12. Conenello GM, Zamarin D, Perrone LA, Tumpey T, Palese P(2007) A single mutation in the PB1-F2 of H5N1 (HK/97) and 1918 influenza A viruses contributes to increased virulence. *PLoS Pathog* **3**: 1414-21.

13. Crescenzo-Chaigne B, Naffakh N, van der Werf S(1999) Comparative analysis of the ability of the polymerase complexes of influenza viruses type A, B and C to assemble into functional RNPs that allow expression and replication of heterotypic model RNA templates in vivo. *Virology* **265**: 342-53.

14. Duan L, Bahl J, Smith GJ *et al.*(2008) The development and genetic diversity of H5N1 influenza virus in China, 1996-2006. *Virology* **380**: 243-54.

15. Dugan VG, Chen R, Spiro DJ *et al.*(2008) The evolutionary genetics and emergence of avian influenza viruses in wild birds. *PLoS Pathog* **4**: e1000076.

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

17. Gabriel G, Dauber B, Wolff T, Planz O, Klenk HD, Stech J(2005) The viral polymerase mediates adaptation of an avian influenza virus to a mammalian host. *Proc Natl Acad Sci U S A* **102**: 18590-5.

18. Garten RJ, Davis CT, Russell CA *et al.*(2009) Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* **325**: 197-201.

19. Guan Y, Peiris JS, Lipatov AS *et al.*(2002) Emergence of multiple genotypes of H5N1 avian influenza viruses in Hong Kong SAR. *Proc Natl Acad Sci U S A* **99**: 8950-5.

20. Guan Y, Peiris M, Kong KF *et al.*(2002) H5N1 Influenza Viruses Isolated from Geese in Southeastern China: Evidence for Genetic Reassortment and Interspecies Transmission to Ducks. *Virology* **292**: 16-23.

21. Guan Y, Poon LL, Cheung CY *et al.*(2004) H5N1 influenza: a protean pandemic threat. *Proc Natl Acad Sci U S A* **101**: 8156-61.

22. Guan Y, Shortridge KF, Krauss S *et al.*(2000) H9N2 influenza viruses possessing H5N1-like internal genomes continue to circulate in poultry in southeastern China. *J Virol* **74**: 9372-80.

23. Guan Y, Shortridge KF, Krauss S, Webster RG(1999) Molecular characterization of H9N2 influenza viruses: were they the donors of the "internal" genes of H5N1 viruses in Hong Kong? *Proc Natl Acad Sci U S A* **96**: 9363-7.

24. Guindon S, Gascuel O(2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696-704.

25. Hatta M, Gao P, Halfmann P, Kawaoka Y(2001) Molecular Basis for High Virulence of Hong Kong H5N1 Influenza A Viruses. *Science* **293**: 1840-2.

26. Hoffmann E, Stech J, Leneva I *et al.*(2000) Characterization of the influenza A virus gene pool in avian species in southern China: was H6N1 a derivative or a precursor of H5N1? *J Virol* **74**: 6309-15.

27. Jackson D, Hossain MJ, Hickman D, Perez DR, Lamb RA(2008) A new influenza virus virulence determinant: the NS1 protein four C-terminal residues modulate pathogenicity. *Proc Natl Acad Sci U S A* **105**: 4381-6.

28. Kawaoka Y, Gorman OT, Ito T *et al.*(1998) Influence of host species on the evolution of the nonstructural (NS) gene of influenza A viruses. *Virus Res* **55**: 143-56.

29. Li KS, Guan Y, Wang J *et al.*(2004) Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature* **430**: 209-13.

30. Li Z, Chen H, Jiao P *et al.*(2005) Molecular basis of replication of duck H5N1 influenza viruses in a mammalian mouse model. *J Virol* **79**: 12058-64.

31. Long JX, Peng DX, Liu YL, Wu YT, Liu XF(2008) Virulence of H5N1 avian influenza virus enhanced by a 15-nucleotide deletion in the viral nonstructural gene. *Virus Genes* **36**: 471-8.

32. Lu G, Rowley T, Garten R, Donis RO(2007) FluGenome: a web tool for genotyping influenza A virus. *Nucleic Acids Res* **35**: W275-W279.

33. Macken CA, Webby RJ, Bruno WJ(2006) Genotype turnover by reassortment of replication complex genes from avian influenza A virus. *J Gen Virol* **87**: 2803-15.

34. Massin P, van der Werf S, Naffakh N(2001) Residue 627 of PB2 is a determinant of cold sensitivity in RNA replication of avian influenza viruses. *J Virol* **75**: 5398-404.

35. Mazur I, Anhlan D, Mitzner D, Wixler L, Schubert U, Ludwig S(2008) The proapoptotic influenza A virus protein PB1-F2 regulates viral polymerase activity by interaction with the PB1 protein. *Cell Microbiol* **10**: 1140-52.

36. McAuley JL, Hornung F, Boyd KL *et al.*(2007) Expression of the 1918 influenza A virus PB1-F2 enhances the pathogenesis of viral and secondary bacterial pneumonia. *Cell Host Microbe* **2**: 240-9.

37. Naffakh N, Massin P, Escriou N, Crescenzo-Chaigne B, van der Werf S(2000) Genetic analysis of the compatibility between polymerase proteins from human and avian strains of influenza A viruses. *J Gen Virol* **5**: 1283-91.

38. Nelson MI, Viboud C, Simonsen L *et al.*(2008) Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. *PLoS Pathog* **4**: e1000012.

39. Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team(2009) Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans. *N Engl J Med.*

40. Obenauer JC, Denson J, Mehta PK *et al.*(2006) Large-Scale Sequence Analysis of Avian Influenza Isolates. *Science* **311**: 1576-80.

41. Palese P, Shaw M (2007) Orthomyxoviridae: The Viruses and Their Replication. In: Knipe DM, Howley PM, Griffin DE, Lamb RA, Martin MA, Roizman B, Straus SE, eds. *Fields Virology*. Philadelphia: Wolters Kluwer, Lippincott Williams & Wilkins, pp. 1647-1689.

42. Salomon R, Franks J, Govorkova EA *et al.*(2006) The polymerase complex genes contribute to the high virulence of the human H5N1 influenza virus isolate A/Vietnam/1203/04. *J Exp Med* **203**: 689-97.

43. Shinya K, Hamm S, Hatta M, Ito H, Ito T, Kawaoka Y(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**: 258-66.

44. Smith GJ, Fan XH, Wang J *et al.*(2006) Emergence and predominance of an H5N1 influenza variant in China. *Proc Natl Acad Sci U S A* **103**: 16936-41.

45. Smith GJ, Naipospos TS, Nguyen TD *et al.*(2006) Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. *Virology* **350**: 258-68.

46. Smith GJ, Vijaykrishna D, Ellis TM *et al.*(2009) Characterization of avian influenza viruses A (H5N1) from wild birds, Hong Kong, 2004-2008. *Emerg Infect Dis* **15**: 402-7.

47. Subbarao EK, London W, Murphy BR(1993) A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. *J Virol* **67**: 1761-4.

48. Subbarao K, Klimov A, Katz J *et al.*(1998) Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness. *Science* **279**: 393-6.

49. Sweet TM, Maassab HF, Herlocher ML(2004) Reverse genetics studies of attenuation of the ca A/AA/6/60 influenza virus: the role of the matrix gene. *Biomed Pharmacother* **58**: 509-15.

50. Van HN, Pappas C, Belser JA *et al.*(2009) Human HA and polymerase subunit PB2 proteins confer transmission of an avian influenza virus through the air. *Proc Natl Acad Sci U S A* **106**: 3366-71.

51. Vijaykrishna D, Bahl J, Riley S *et al.*(2008) Evolutionary dynamics and emergence of panzootic H5N1 influenza viruses. *PLoS Pathog* **4**: e1000161.

52. Watanabe T, Watanabe S, Shinya K, Kim JH, Hatta M, Kawaoka Y(2009) Viral RNA polymerase complex promotes optimal growth of 1918 virus in the lower respiratory tract of ferrets. *Proc Natl Acad Sci U S A* **106**: 588-92.

53. Webster RG, Guan Y, Peiris M *et al.*(2002) Characterization of H5N1 influenza viruses that continue to circulate in geese in southeastern China. *J Virol* **76**: 118-26.

54. Wright PF, Neumann G, Kawaoka Y (2007) Orthomyxoviruses. In: Knipe DM, Howley PM, Griffin DE, Lamb RA, Martin MA, Roizman B, Straus SE, eds. *Fields Virology*. Philadelphia: Wolters Kluwer, Lippincott Willams & Wilkins, pp. 1691-1740.

55. Xu KM, Smith GJ, Bahl J *et al.*(2007) The genesis and evolution of H9N2 influenza viruses in poultry from southern China, 2000 to 2005. *J Virol* **81**: 10389-401.

56. Xu X, Subbarao, Cox NJ, Guo Y(1999) Genetic characterization of the pathogenic influenza A/Goose/Guangdong/1/96 (H5N1) virus: similarity of its hemagglutinin gene to those of H5N1 viruses from the 1997 outbreaks in Hong Kong. *Virology* **261**: 15-9.

57. Zamarin D, Garcia-Sastre A, Xiao X, Wang R, Palese P(2005) Influenza virus PB1-F2 protein induces cell death through mitochondrial ANT3 and VDAC1. *PLoS Pathog* **1**: e4.

SUPPLEMENTARY MATERIAL

(I) Two-time test

The *two-time test* classifies the reassortant status of viruses in time period T2 relative to distinct genotypes circulating in the preceding time period, T1. The two-time test is performed in two steps: first, categorize the genotypic composition of the circulating viral population in T1 by constructing *reference trees*; second, ascertain the reassortment status of T2 viruses by measuring bootstrap support for their placement on the segment-specific reference trees. A T2 virus is declared a *reassortant* if, when added to segment-specific reference trees, it associates with > 70% bootstrap support with different viruses in at least two of the trees[1]. Otherwise, the T2 virus is declared a *non-reassortant.* Strictly speaking, the terms "reassortant" and "non-reassortant" are only relative: the classification of reassortment status of a T2 virus is *relative* to the genotypes in the previous time period, T1. The results of our study (4) and studies of others (1-3,5) suggest that the genotype population is constantly turning over due to the reassortment process.

(a) Choice of taxa for reference trees: Reference trees are required (i) to include as many as possible of the distinguishable genotypes of T1 viruses, in order to be able to accurately classify the reassortant status of T2 viruses relative to potential donor viruses, (ii) to have high bootstrap support for their topological features, so that the classification of T2 viruses can be unequivocal.

(b) Inference of reference trees: We begin with a set of phylogenetic trees, one for each segment; each tree includes exactly the same set of candidate reference viruses. We identify *distinguishable* genotypes and select one representative of each distinguishable genotype as follows. Starting from the tips of the segment-specific phylogenies, we identified clusters of viruses colocated in every segment-specific phylogenetic tree with high bootstrap support. We

[1] We have assumed that the data set does not include recombinant segments. If recombinants are present in our reference trees, our results still hold except that in some instances declared reassortants are probably recombinants.

select a single member of the cluster to represent the cluster genotype. This grouping and selecting reduces the T1 dataset to a subset of viruses, each representing a distinguishable genotype. We describe this complete subset of distinguished genotypes as the *reference set.* The phylogenies of this subset become our "draft" reference trees.

In order to have confidence in the classification of the reassortant status of T2 viruses relative to reference trees, reference trees needed a high degree of topological certainty. In our experience, it is typical that some of the viruses in the draft reference trees attach at deeper nodes (i.e., not at the tips) with low (< 70%) bootstrap support in one or more of the segment-specific draft reference trees; these viruses are removed from the draft reference trees to produce the final reference trees, but - importantly - are retained in the reference set to be used as candidate donor segments as described below (S.Ic).

(c) Genotyping T2 (test) viruses: When we characterize the genotypes of T2 viruses, we first look for donor segments among the viruses of the reference tree. To do this, we add a single test virus to the reference trees. We assess the bootstrap support for a test virus associating with a single reference virus. If the test virus has at least 70% bootstrap support for being the sister to a reference virus, then we say that this reference virus represents the lineage from which the segment of the test virus was derived. If no donor is apparent (i.e. if the attachment point is basal to multiple reference tree viruses), we search the viruses that are included in the reference set, but not in the reference tree, for the donor. To do this, we add a single virus from the reference set that is not already in the reference trees to the reference trees. We then assess the bootstrap support for a test virus associating with a single virus in this augmented reference tree. We repeat this process of single-virus augmentation and assessment of bootstrap support for association of a test virus with a single virus in the augmented reference tree until all viruses from the reference set that are not in the original reference tree have been added, one at a time, to the reference tree. If no virus in an augmented reference tree meets the criterion for being a representative of the donor lineage,

then we say that the segment of the test virus cannot be classified, and therefore the test virus cannot be genotyped. When test viruses attach to a reference tree at deeper nodes, our interpretation is that the donor lineage is not in our dataset and the segment is from a novel source.

Inclusion of the viruses in the reference set that are not included in the original reference tree thus reduces the chances of mislabelling a T2 virus as a reassortant simply because its genotype, while present in the T1 population, is not in the reference tree.

In our previous study, we demonstrated that the topologies of, and bootstrap support for, the reference trees for a variety of methods of phylogenetic inference were the same and bootstrap support was similar for every node; therefore, the assessment of reassortment was robust to the method of phylogenetic inference (4).

(d) Modifications for testing closely related highly pathogenic avian H5N1 viruses: The close relationships among avian H5N1 viruses sometimes led to a situation in which two H5N1 viruses were readily distinguishable in the two-time test for some segments, but not distinguishable for other segments. An extreme example is the pair of viruses in the reference trees for the first time period, A/goose/Guangdong/1/96 and A/duck/Guangxi/07/99 (Fig. 1). These two viruses have distinct genotypes; our algorithm for inferring reference trees led to both being included in the reference trees. The NS segments were substantially different, as they came from different alleles. The five other segments were not identical, but were not distinguishable phylogenetically in our two-time test. Specifically, these two viruses were sisters in all segment-specific reference trees except that for NS. When we added a test virus to the reference tree, among those test viruses that were sisters with either of A/goose/Guangdong/1/96 and A/duck/Guangxi/07/99, almost all had bootstrap values below our threshold of 70% to one or both of A/goose/Guangdong/1/96 and A/duck/Guangxi/07/99, but high bootstrap support (generally 90 - 100%) for being in a clade with these two viruses. Thus, although the segments of A/goose/Guangdong/1/96 and A/duck/Guangxi/07/99 were clearly not

identical, they could not be distinguished phylogenetically. In this study, when a segment of distinct viruses could not be distinguished phylogenetically, we represented the segment of both viruses with the same line style and color.

(e) Identification of sub-genotypes: This divergence of genotype Z into sub-genotypes became apparent in the derivation of the reference trees for the third time period. Three candidate reference viruses, all classified as genotype Z in the second time period, were indistinguishable phylogenetically for the NP and M segments, but represented distinct clusters of candidate reference viruses for PB2, PB1, PA, and NS segments. For this reason, we retained all three in the reference trees even though no obvious reassortment events distinguished them. If these three viruses were not in fact phylogenetically distinct, then test viruses would have equivocal associations with them in the third period test process, as discussed in (S.Id) above. However, when testing viruses from 2005 - 2008, we found that these three reference viruses were phylogenetically distinguishable: in the subset of test viruses that associated most closely with any of these three viruses, with extremely rare exceptions, they had high bootstrap support for association with at most one of the three (PB2, PB1, PA) or two of the three (NS) segments.

(II) Supplementary Table:

Supplementary Table 1: Relationships between genotype classification of this publication and alternative classification systems.

Virus	Year	HA Clade	Genotype	
			This study (see Figs. 1-3)	Alternative classification systems (1,5,6)
A/duck/Guangxi/35/01	2001	3	9	B
A/Chicken/Hong Kong/YU562/01	2001	3	12	B
A/duck/Fujian/17/01	2001	3	12	B
A/duck/Guangdong/22/02	2002	4	19	B
A/goose/Fujian/bb/03	2003	4	19	B
A/curlew/Shandong/61/04	2004	6	21	B
A/chicken/Jiangsu/cz1/02	2002	NC[1]	23	B
A/goose/Shantou/157/02	2002	NC	24	B
A/duck/Guangxi/22/01	2001	3	12	B1
A/duck/Shantou/1930/01	2001	0	12	B2
A/partridge/Shantou/478/02	2002	4	ND[2]	B3
A/chicken/Shantou/904/01	2001	3	12	C
A/duck/Guangdong/01/01	2001	3	12	C
A/duck/Hong Kong/2986.1/00	2000	3	12	C
A/duck/Shanghai/13/01	2001	3	12	C
A/Chicken/Hong Kong/FY150/01	2001	3	13	D
A/duck/Guangxi/13/04	2004	2.4	17	G
A/common magpie/Hong Kong/2125/06	2006	2.3.4	32	G
A/common magpie/Hong Kong/2256/06	2006	2.3.4	32	G
A/common magpie/Hong Kong/3033/06	2006	2.3.4	32	G
A/house crow/Hong Kong/2648/06	2006	2.3.4	32	G
A/house crow/Hong Kong/2858/06	2006	2.3.4	32	G
A/Japanese white-eye/Hong Kong/1038/06	2006	2.3.4	32	G
A/large-billed crow/Hong Kong/2512/06	2006	2.3.4	32	G
A/munia/Hong Kong/2454/06	2006	2.3.4	32	G
A/white-backed munia/Hong Kong/2469/06	2006	2.3.4	32	G
A/duck/Guangxi/3548/2005	2005	2.3.2	34	G
A/duck/Hunan/1265/2005	2005	2.3.2	34	G
A/duck/Yunnan/4400/05	2005	2.3.2	34	G
A/duck/Yunnan/6607/2005	2005	2.3.2	34	G
A/goose/Guangxi/3316/2005	2005	2.3.2	34	G
A/goose/Guangxi/345/05	2005	2.3.2	34	G
A/duck/Guangxi/89/2006	2006	2.3.2	34	G
A/pheasant/Shantou/2239/2006	2006	2.3.2	34	G
A/goose/Guiyang/337/06	2006	4	ND	G
A/chicken/Hebei/718/01	2001	3	10	Hubei-like
A/mallard/Guangxi/wt/04	2004	9	ND	Hubei-like
A/tree sparrow/Henan/1/04	2004	6	ND	Hubei-like
A/duck/Fujian/19/00	2000	0	4	T1

A/duck/Guangdong/40/00	2000	0	8	T2
A/chicken/Jilin/9/04	2004	NC	ND	T5
A/wild duck/Hunan/211/05	2005	2.3.4	ND	T6
A/blackbird/Hunan/1/04	2004	6	16	V
A/Ck/ST/4231/03	2003	2.5	16	V
A/common magpie/Hong Kong/645/06	2006	2.3.4	33	V
A/crested myna/Hong Kong/540/06	2006	2.3.4	33	V
A/duck/Hunan/856/06	2006	2.3.4	33	V
A/chicken/Hong Kong/282/06	2006	2.3.4	35	V
A/chicken/Hong Kong/947/06	2006	2.3.4	35	V
A/duck/Guangxi/150/06	2006	2.3.4	35	V
A/little egret/Hong Kong/718/06	2006	2.3.4	35	V
A/duck/Fujian/668/06	2006	2.3.4	36	V
A/migratory duck/Jiangxi/2136/05	2005	2.2	37	V
A/goose/Shantou/239/06	2006	2.5	ND	V2
A/duck/Zhejiang/52/00	2000	5	7	W
A/goose/Guangxi/1097/04	2004	5	18	W
A/quail/Shantou/3846/02	2002	5	18	W
A/chicken/Shantou/2535/01	2001	0	11	W1
A/Duck/Anyang/AVL-1/01	2001	0	3	X0
A/duck/Shandong/093/04	2004	0	3	X0
A/duck/Shanghai/38/01	2001	0	5	X0
A/duck/Shantou/4912/01	2001	0	5	X0
A/goose/Shantou/5456/01	2001	0	5	X0
A/chicken/Shantou/28/02	2002	0	27	X0
A/SCk/Hong Kong/YU100/02	2002	0	ND	X3
A/duck/Shanghai/35/02	2002	0	ND	X4
A/duck/Shanghai/37/02	2002	0	29	X5
A/duck/Guangxi/53/02	2002	0	25	X6
A/duck/Fujian/13/02	2002	0	30	X7
A/chicken/Hebei/108/02	2002	0	3	X8
A/chicken/Shantou/9/02	2002	0	26	X9
A/chicken/Shantou/5738/01	2001	0	6	X10
A/chicken/Shantou/5746/01	2001	0	6	X10
A/chicken/Hunan/23/02	2002	NC	14	Z
A/chicken/Shantou/4059/02	2002	NC	14	Z
A/chicken/Yunnan/1215/02	2002	1	14	Z
A/chicken/Yunnan/1252/03	2003	1	14	Z
A/chicken/Yunnan/1628/03	2003	NC	14	Z
A/duck/Guangxi/50/01	2001	9	14	Z
A/duck/Hunan/1340/02	2002	NC	14	Z
A/duck/Hunan/1386/03	2003	1	14	Z
A/duck/Hunan/300/03	2003	NC	14	Z
A/duck/Hunan/795/02	2002	NC	14	Z
A/duck/Shanghai/xj/02	2002	NC	14	Z
A/duck/Yunnan/119/03	2003	1	14	Z
A/duck/Yunnan/971/02	2002	NC	14	Z
A/partridge/Shantou/1075/2002	2002	9	14	Z
A/peregrine falcon/Hong Kong/D0028/04	2004	9	14	Z

A/pheasant/Shantou/40/03	2003	1	14	Z
A/Bar-headed Goose/Qinghai/65/05	2005	2.2	Z.1	Z
A/Guinea fowl/Shantou/1341/06	2006	2.2	Z.1	Z
A/turkey/Turkey/1/05	2005	2.2.1	Z.1	Z
A/chicken/Hunan/999/05	2005	2.3.1	Z.3	Z
A/babbler/Fujian/320/04	2004	6	14	Z+

[1]NC: HA clade not classified; [2]ND: Genotype not designated.

References

1. Duan L, Bahl J, Smith GJ *et al.*(2008) The development and genetic diversity of H5N1 influenza virus in China, 1996-2006. *Virology* **380**: 243-54.

2. Dugan VG, Chen R, Spiro DJ *et al.*(2008) The evolutionary genetics and emergence of avian influenza viruses in wild birds. *PLoS Pathog* **4**: e1000076.

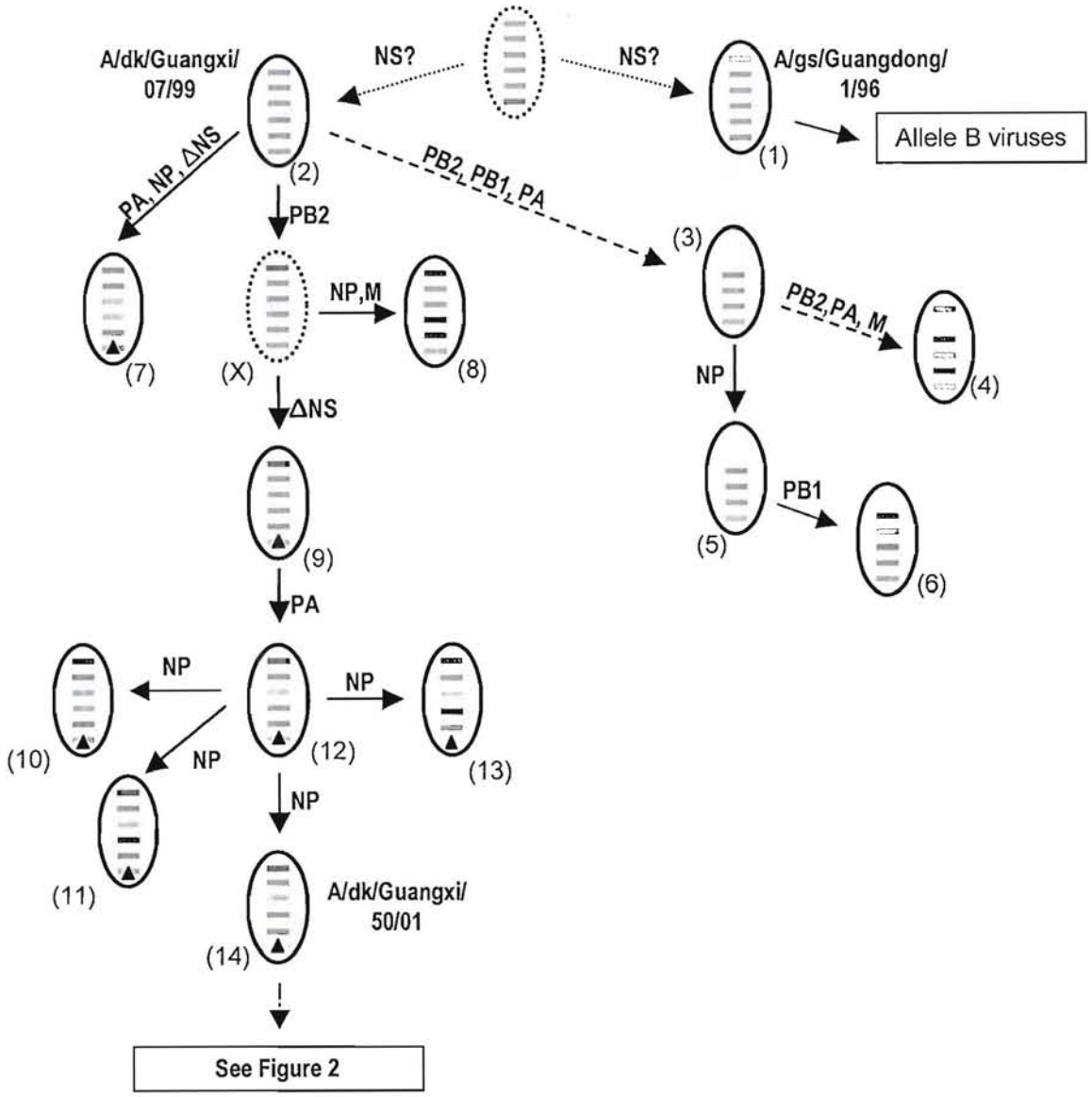
3. Li KS, Guan Y, Wang J *et al.*(2004) Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature* **430**: 209-13.

4. Macken CA, Webby RJ, Bruno WJ(2006) Genotype turnover by reassortment of replication complex genes from avian influenza A virus. *J Gen Virol* **87**: 2803-15.

5. Smith GJ, Fan XH, Wang J *et al.*(2006) Emergence and predominance of an H5N1 influenza variant in China. *Proc Natl Acad Sci U S A* **103**: 16936-41.

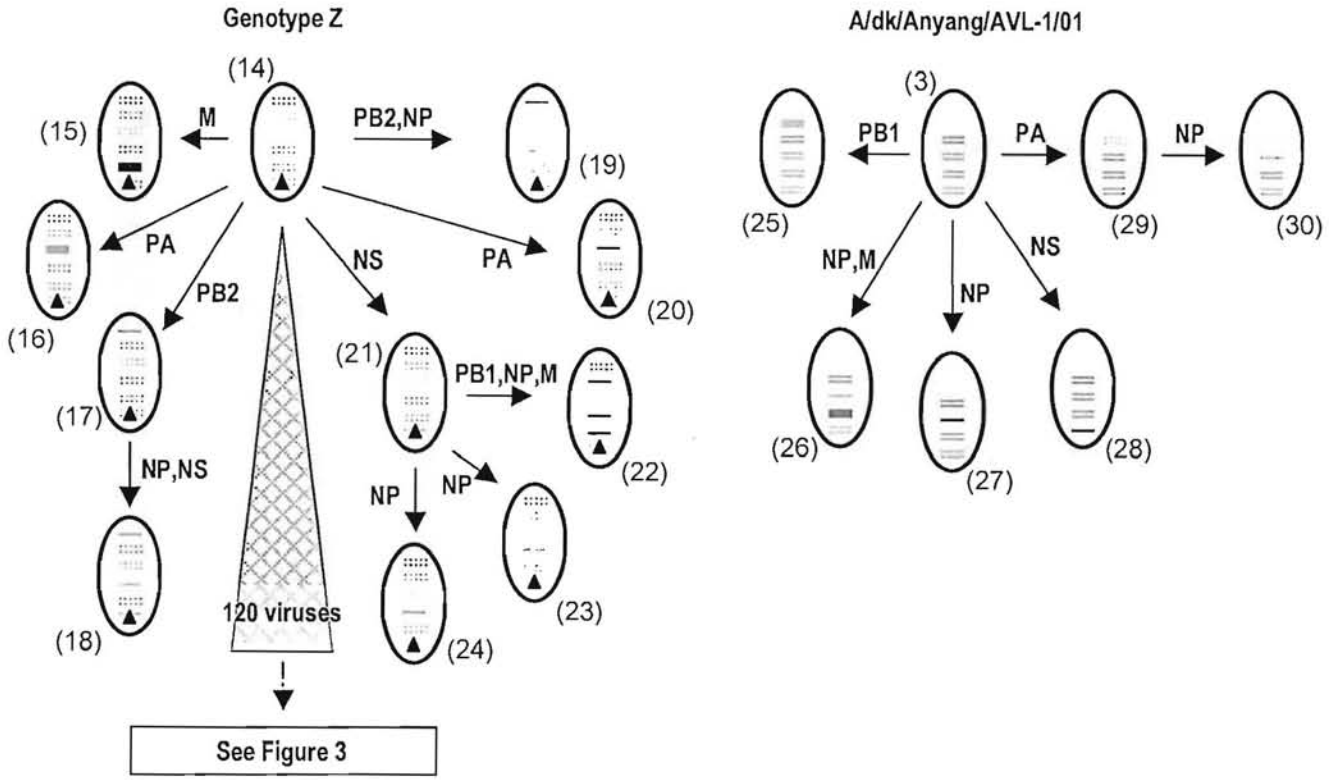
6. Smith GJ, Vijaykrishna D, Ellis TM *et al.*(2009) Characterization of avian influenza viruses A (H5N1) from wild birds, Hong Kong, 2004-2008. *Emerg Infect Dis* **15**: 402-7.

Fig. 1



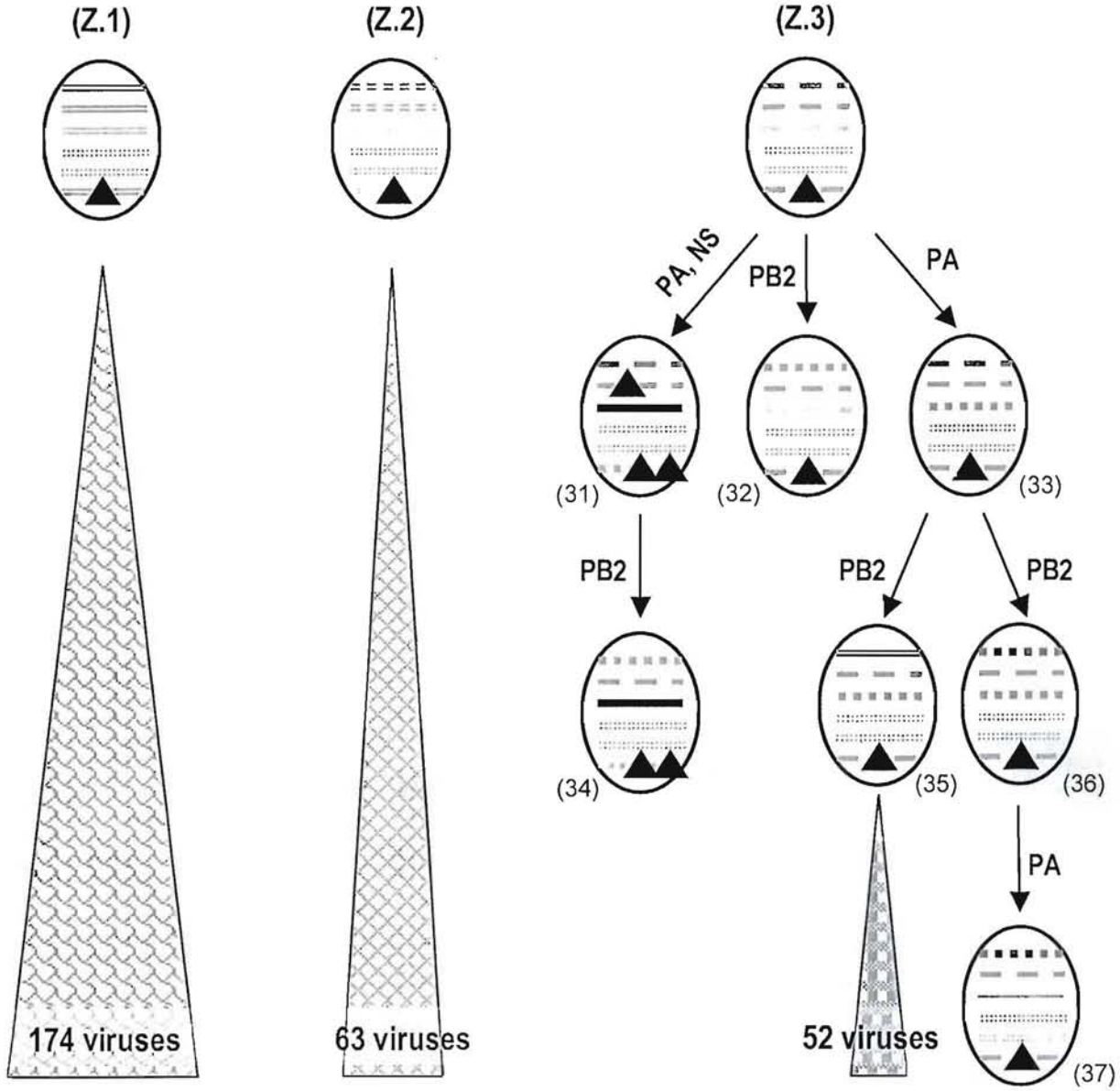
Genotype	Representative virus	History of generation	Viruses per genotype	Clade
(1)	A/gs/Guangdong/1/96		1	0
(2)	A/dk/Guangxi/07/99	Novel Allele A NS	1	0
(3)	A/dk/Anyang/AVL-1/01	A/ostrich/South Africa/9508103/95(H9N2)-like PB2,PB1 A/chicken/Shanghai/F/98(H9N2)-like PA	1	0
(4)	A/dk/Fujian/19/00	A/chicken/Shanghai/F/98(H9N2)-like PB2 A/chicken/Hubei/wm/97(H5N1)-like PA Unknown M	1	0
(5)	A/dk/Shanghai/38/01	A/chicken/Shanghai/F/98(H9N2)-like NP	7	0
(6)	A/ck/Shantou/5738/01	A/quail/Shanghai/8/96(H9N2)-like PB1	2	0
(7)	A/dk/Zhejiang/52/00	A/duck/Hong Kong/Y439/97(H9N2)-like PA,NP Internal NS1 deletion	1	5
(X)	Missing link	A/duck/Nanchang/1749/92(H11N2)-like PB2		
(8)	A/dk/Guangdong/40/00	Unknown NP, M	1	0
(9)	A/dk/Guangxi/35/01	Internal NS1 deletion	2	3
(10)	A/ck/Hebei/718/01	A/African Starling/England-Q/938/1979(H7N1)-like NP	1	3
(11)	A/ck/Shantou/2535/01	Unknown NP	1	0
(12)	A/dk/Fujian/17/01	A/duck/Hong Kong/Y439/97(H9N2)-like PA	13	3
(13)	A/ck/HK/FY150/01	A/quail/Shanghai/8/96(H9N2)-like NP	1	3
(14)	A/dk/Guangxi/50/01	A/chicken/Shanghai/F/98(H9N2)-like NP	2	3,9

Fig. 2



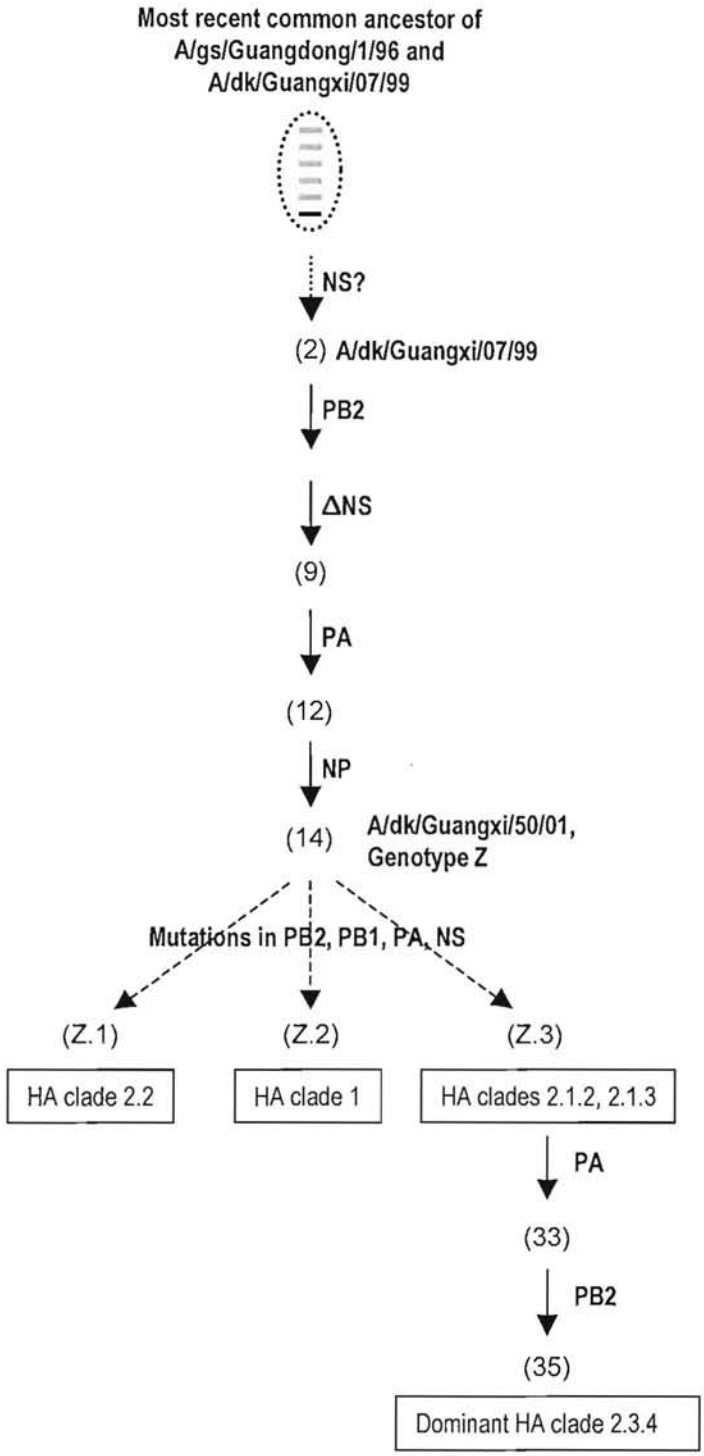
Genotype	Representative virus	History of generation	Viruses per genotype	Clade
(3)	A/dk/Anyang/AVL-1/01	Evolved from A/dk/Anyang/AVL-1/01 by mutations in all genes	5	0
(14)	A/dk/Guangxi/50/01	Evolved from A/dk/Guangxi/50/01 by mutations in all genes	120	1+Others
(15)	A/ck/Hubei/wo/03	A/pheasant/Hong Kong/FY294/00(H6N1)-like M	1	7
(16)	A/ck/Korea/ES/03	A/ck/Taiwan/165/99 (H6N1)-like PA	11	2.5
(17)	A/ck/Guangxi/12/04	A/dk/Zhejiang/52/00 (H5N1)-like PB2	2	2.4
(18)	A/quail/Shantou/3846/02	A/dk/Zhejiang/52/00(H5N1)-like NP	10	5
		A/dk/Zhejiang/52/00(H5N1)-like NS		
(19)	A/dk/Shantou/700/02	A/dk/Fujian/17/2001(H5N1)-like PB2	3	4
		A/dk/Guangxi/35/01(H5N1)-like NP		
(20)	A/wild duck/Shanghai/59/04	Unknown PA segment	1	2.5
(21)	A/curlew/Shandong/61/04	A/dk/Zhejiang/52/00(H5N1)-like NS	1	6
(22)	A/ck/Jilin/9/04	Unknown PB1, NP, and M segments	1	NC
(23)	A/ck/Jiangsu/cz1/02	A/dk/Guangxi/35/01(H5N1)-like NP	1	NC
(24)	A/gs/Shantou/157/02	A/dk/Shanghai/38/01(H5N1)-like NP	1	NC
(25)	A/dk/Guangxi/52/02	A/gs/Guangdong/1/96(H5N1)-like PB1	1	0
(26)	A/ck/Shantou/9/02	A/dk/Guangxi/35/01(H5N1)-like NP	1	0
		A/pheasant/Hong Kong/FY294/00(H6N1)-like M		
(27)	A/ck/Shantou/28/02	Unknown NP	1	0
(28)	A/dk/Fujian/01/02	Unknown NS	13	0
(29)	A/dk/Shanghai/37/02	A/dk/Guangi/50/01(H5N1)-like PA	1	0
(30)	A/dk/Fujian/13/02	A/dk/Guangxi/35/01(H5N1)-like NP	2	0

Fig. 3



Genotype	Representative virus	History of generation	Viruses per genotype	Clade
(Z.1)	A/peregrine falcon/HK/D0028/04	Evolved from 'Z' by mutations in PB2, PB1, PA, NS	174	2.2
(Z.2)	A/ck/Yunnan/1252/03	Evolved from 'Z' by mutations in PB2, PB1, PA, NS	63	1
(Z.3)	A/dk/Hunan/782/03	Evolved from 'Z' by mutations in PB2, PB1, PA, NS	13	2.1.2, 2.1.3
(31)	A/ck/Guiyang/3055/05	A/dk/Guangxi/1311/04 (H5N1)-like NS Unknown PA (likely of genotype Z origin)	7	2.3.4
(32)	A/common magpie/HK/2125/06	A/dk/Guangxi/1311/04 (H5N1)-like PB2	9	2.3.4
(33)	A/dk/Fujian/11094/05	A/ck/Guangdong/174/04 (H5N1)-like PA	30	2.3.4
(34)	A/ck/Guangxi/3154/05	A/dk/Guangxi/1311/04(H5N1)-like PB2	38	2.3.2
(35)	A/dk/Guangxi/5165/05	Sub-lineage Z.1-like PB2	52	2.3.4
(36)	A/dk/Fujian/668/06	A/ck/Guangdong/174/04(H5N1)-like PB2	4	2.3.4
(37)	A/dk/Yunnan/5133/05	A/blackbird/Hunan/1/04(H5N1)-like PA	6	2.2,7

Fig. 4


Most recent common ancestor of
A/gs/Guangdong/1/96 and
A/dk/Guangxi/07/99
NS?
(2) A/dk/Guangxi/07/99
PB2
ΔNS
(9)
PA
(12)
NP
(14)
A/dk/Guangxi/50/01,
Genotype Z
Mutations in PB2, PB1, PA, NS
(Z.1)
(Z.2)
(Z.3)
HA clade 2.2
HA clade 1
HA clades 2.1.2, 2.1.3
PA
(33)
PB2
(35)
Dominant HA clade 2.3.4