Evolution of highly pathogenic avian H5N1 influenza viruses and the emergence of dominant variants

Gabrielle Neumann,1 Margaret A. Green2 and Catherine A. Macken2

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

Highly pathogenic avian H5N1 viruses have circulated in South-east Asia for more than a decade and have now spread to more than 60 countries. The evolution of these viruses is characterized by frequent reassortment of the so-called ‘internal’ genes, creating novel genotypes. Additionally, over time, the surface glycoprotein, haemagglutinin (HA), which is the primary target of the adaptive immune response, has evolved by point mutation into 20 genetically and potentially antigenically distinct clades. To investigate the evolution of avian H5N1 influenza viruses, we undertook a high-resolution analysis of the reassortment of internal genes and evolution of HA of 651 avian H5N1 viruses from 2000 to 2008. Our analysis suggested: (i) all current H5N1 genotypes were derived from a single, clearly defined sequence of initial reassortment events; (ii) reassortment of just three of the internal genes had the most importance in avian H5N1 virus evolution; (iii) HA and the constellation of internal genes may be jointly important in the emergence of dominant variants. Further, our analysis led to the identification of evolutionarily significant molecular changes in the internal genes that may be significant for the emergence of these dominant variants.

INTRODUCTION

Outbreaks of highly pathogenic avian influenza viruses of the H5N1 subtype were first reported in 1996/1997 in southern China and Hong Kong (de Jong et al., 1997; Subbarao et al., 1998; Claas et al., 1998a, b; Xu et al., 1999). 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 not yet understood. Further, their substantial antigenic variation (WHO, 2008) makes the development of vaccines problematic.

The genomic material of influenza A viruses is divided among eight segments, each coding for one or two proteins (Palese & Shaw, 2007). The surface proteins haemagglutinin (HA) and neuraminidase (NA), each encoded by a single segment, define the antigenic properties of the virus. HA is the primary target of the adaptive immune response. Currently, 16 HA (H1–16) and nine NA subtypes (N1–9) are recognized (Wright et al., 2007). The remaining six (‘internal’) segments code for proteins with functions in viral replication, assembly and budding, and interference with host innate immunity (Palese & Shaw, 2007).

The segmentation of the influenza viral genome enables the process of reassortment, when a single host cell is infected by at least two distinct viruses. Novel progeny viruses can be generated when 1–7 segments of one virus are replaced by the corresponding segments of another virus. Reassortment is critical for influenza virus evolution; the novel influenza viruses involved in the past three pandemics were all reassortants (Wright et al., 2007; Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team et al., 2009; Garten et al., 2009; Neumann et al., 2009). Reassortment also contributes to ‘daily’ virus evolution, for example, among avian influenza viruses in their wild aquatic hosts (Dugan et al., 2008), and among human influenza A viruses (Nelson et al., 2008).

The highly pathogenic avian H5N1 influenza viruses that emerged in 1996/1997 were reassortants with an A/goose/Guangdong/1/96 (H5N1)-like HA gene (Xu et al., 1999); other segments were drawn from genetic lineages not native to A/goose/Guangdong/1/96. These non-native lineages are most closely represented in the current database by H5N1, H6N1 and H9N2 viruses (Guan et al., 1999; Hoffmann et al., 2000; Cheung et al., 2007). Since this initial outbreak, avian H5N1 viruses evolved substantially due to reassortment with viruses of the H5N1, H7N1, H9N2 and H11N2 subtypes (Guan et al., 2002a, 2004; Li et al., 2004; Smith et al., 2006a; Chen et al., 2006a, b; Duan et al., 2008) until, in 2002, a predominant genotype (‘Z’)
was recognized (Guan et al., 2004; Li et al., 2004; Smith et al., 2006a). Genotype Z viruses have, to date, persisted and spread geographically throughout Asia, Africa and Europe. Non-genotype Z constellations have been observed, but considerably less frequently (Duan et al., 2008). Meanwhile, the HA proteins of avian H5N1 viruses, while derived from a single lineage, have evolved by point mutation into distinguishable clades, numbered 0–9 (with some second- and third-order subdivisions) according to a unified nomenclature system (WHO, 2008).

A number of published studies on the reassortment of H5N1 viruses have identified and characterized a large number of distinct genotypes (Guan et al., 2002a, 2004; Li et al., 2004; Smith et al., 2006a; Chen et al., 2006a, b; Duan et al., 2008). Also, the substantial antigenic variation of the HA protein has been extensively studied (WHO Global Influenza Program Surveillance Network, 2005; WHO, 2008; Wan et al., 2008; Smith et al., 2009), with obvious relevance to the selection of candidate vaccine strains. However, it is not clear if these distinct processes of reassortment of the internal genes and antigenic evolution of HA by point mutation interact in the emergence of dominant variants of avian H5N1 viruses. The aim of this present study is to investigate the roles of these dual evolutionary processes in the evolution of avian H5N1 viruses from the period 2000 to 2008 during which the three currently dominant antigenic variants of HA have emerged.

RESULTS

Both reassortment and point mutation have contributed significantly to the evolution of avian H5N1 viruses. To explore the contributions of these two distinct evolutionary processes to the evolution of avian H5N1 viruses, we carried out a high-resolution analysis of an extensive dataset of avian H5N1 viruses from 2000 to 2008, the period within which the currently dominant HA clades emerged. We analysed avian H5N1 viruses for which complete, or almost complete, genomic sequences are available. Human isolates were excluded to avoid the distortion of datasets with mutations that may reflect adaptation to mammalian species. In total, 651 avian H5N1 viruses were genotyped.

First, we genotyped the internal genes using our two-time test (Macken et al., 2006). Although the two-time test for genotyping is more difficult to implement than genotyping using BLAST (Lu et al., 2007), the results are more precise; this extra precision was important for inferring a detailed model of evolution by reassortment. The most parsimonious relationship among the genotypes is represented in the evolutionary paths of Figs 1–3. After genotyping, we identified the HA clade [determined from the unified nomenclature system (WHO, 2008) and from our analyses] of viruses within each genotype. HA clades associated with each genotype are listed in the figure legends.

The two-time test acts like a sliding window in which earlier viruses are viewed as 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. Candidate reference viruses were avian influenza A viruses selected from the period up to the test period; these viruses reflect all serotypes (including H5N1) and include isolates from Asia, Africa and Europe, as these viruses could plausibly donate genetic material to test viruses. We analysed our dataset from 2000 to 2008 in three phases. The first phase included test viruses from 2000 to 2001; frequent reassortment has been reported between 1999 and 2001 (Guan et al., 2002a, b; Li et al., 2004; Chen et al., 2006b; Duan et al., 2008). The second phase included test viruses from 2002 to 2004; this is the period during which the dominance of a particular genotype (‘Z’) was consolidated in South-east Asia. The third phase, 2005–2008, is one of extensive geographical spread of avian H5N1 viruses.

Genotypes of avian H5N1 viruses from 2000 to 2001

The 'test' virus set contained 34 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 %. Two NP and two M gene segments could not be assigned a lineage; bootstrap support for association of these segments with any of the reference viruses fell substantially below the acceptable threshold. We believe this low support was because the donor of the reassorted segments was not present in the database and hence could not be included in the reference set. (Trees for these test viruses are given in Supplementary Material, available in JGV Online.)

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) represented 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 gene (Xu et al., 1999), while the 1997 Hong Kong H5N1 viruses and almost all contemporary H5N1 viruses possess an allele A NS gene (Guan et al., 1999). The common progenitor virus (shown in dotted lines in Fig. 1) is not known, although the chronology of events suggests it is more likely to have an allele B rather than an allele A NS gene.

As described by others (Guan et al., 2002a, b; Li et al., 2004; Chen et al., 2006b; Duan et al., 2008), 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. These novel genotypes were replaced rapidly. Notably, reassortment typically involved replacement of a single segment at a time. Four genotypes represented in
Fig. 1 differed by more than one segment from all other genotypes. This suggested that intermediate viruses were 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 genes) of avian influenza A viruses (Macken et al., 2006), reassorting segments were usually derived from different donor lineages. This behaviour has also been noted implicitly or explicitly by others (Li et al., 2004; Chen et al., 2006b; Dugan et al., 2008; Duan et al., 2008).

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. (2008) 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 have a full-length NA gene; A/duck/Guangxi/50/2001, and – with rare exceptions – more recent avian H5N1, has a 20 aa deletion in the stalk region of the NA gene (Guan et al., 2002a).

Noticeable in early H5N1 virus evolution are frequent reassortment events of the PB2, PB1, PA, and NP genes (which form the viral replication machinery). Another noticeable event in early H5N1 virus evolution is a 5 aa deletion in the NS1 protein that is characteristic of all recent avian H5N1 viruses (Guan et al., 2002a). In our maximum-parsimony model (Fig. 1), the same 5 aa 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. The uncertainty in these early events, when data are sparse, 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/Guangxi/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 from the dataset for this 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 (Duan et al., 2008). These viruses underwent several reassortment events in the period from 2001 to 2002 (genotypes 4–6 in Fig. 1 and genotypes 25–30 in Fig. 2); genotypes 6, 25, 26, 29 and 30 have been described as the transient genotypes X5–7 and X9–10 (Duan et al., 2008). A/ duck/Anyang/AVL-1/01-like viruses were present in 2001–2004, but have not been detected since.

All viruses on the pathway involving A/duck/Anyang/AVL-1/01 had clade 0 HA. By contrast, the HA clade changed several times along the pathway, leading to genotype 14. The two prototype genotype Z (= genotype 14) viruses possessed an HA of clade 3 (A/chicken/Hong Kong/NT873.3/01) or of the genetically distinct clade 9 (A/duck/Guangxi/50/01). Later genotype Z viruses predominantly possessed an HA of clade 1 (see next section and Fig. 2). Currently, it is not known if the substitutions of the HA clade were critical for the evolution of avian H5N1 influenza viruses.

**Genotypes of avian H5N1 viruses from 2002 to 2004**

Our model of evolution for this time period is described in Fig. 2 (note that a model with genotype 21 followed by genotype 18 is equally parsimonious). 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).

The majority of viruses from this test period (127 of 183) were genotype Z viruses, attesting to the dominance of this genotype during the 2002–2004 time period. Genotype Z viruses evolved by point mutations from the founding genotype 14; this evolution is depicted by a line style for genotype 14 that differs between Figs 1 and 2. Genotypes 14 and Z are distinguished by between 5 and 10 conserved nucleotide changes in each of the PB2, PB1, PA and NS gene segments, with no distinguishing changes in the NP and M segments. As in the 2000–2001 time period, most reassortment events involved the polymerase and NP genes. However, in contrast to the previous period, most donor viruses were of the H5N1 subtype. 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.

Notably, eight different HA clades (1, 2.4, 6, 8, 9, 2.1.1, 2.3.1 and 2.3.2) were represented among the 127 genotype Z viruses, with the largest representation (62) from clade 1 HA. Together with the finding that the earliest known genotype Z viruses possessed HA proteins of clades 3 and 9, the substantial diversification of HA in this phase of H5N1 virus evolution may reflect ecological pressure and/or adaptation of HA to the products of internal genes.

Occasional reassortment with H6N1 viruses (Fig. 2), and other unclassifiable and presumably novel viruses (not shown for the sake of clarity) was detected. These viruses provide evidence of the continued introduction of novel genetic material into the H5N1 population, an observation also made by others (Chen et al., 2004; Li et al., 2004; Duan et al., 2008; Vijaykrishna et al., 2008). None of these novel genotypes had descendants in our dataset.

**Genotypes of avian H5N1 viruses from 2005 to 2008**

Our model of evolution for this time period is described in Fig. 3. Our high-resolution genotyping revealed the
Fig. 1. Model of the evolution of avian H5N1 influenza viruses from 2000 to 2001. Depicted are the ‘internal’ genes in the order PB2, PB1, PA, NP, M and NS (six bars, from top to bottom). The bar colour 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. Colour changes indicate reassortment events. ‘ANS’ 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).

divergence of genotype Z viruses into three distinguishable subgenotypes, referred to as Z.1–Z.3 (Fig. 3; see Supplementary Material for details of ascertaining subgenotypes). Subgenotype Z.1 (with the representative virus A/peregrine falcon/Hong Kong/D0028/04) included the Qinghai-Lake viruses and their descendants, which have spread into Europe and Africa. Subgenotype Z.2 (with the representative virus A/chicken/Yunnan/1252/03) was composed of avian H5N1 viruses isolated in Thailand and Vietnam. Subgenotype Z.3 (with the representative virus A/duck/Hunan/782/03) was composed of 31 avian H5N1 viruses, most of which were isolated from chickens in Indonesia.

The three subgenotypes differed from each other and from genotype Z by phylogenetically distinguishable PB2, PB1 and PA genes; some mutations that distinguished these genes resulted in amino acid replacements (in Fig. 3, the diversification into subgenotypes is represented by different line styles of the same colour). In contrast, no such diversification existed for the NP and M segments. The NS segment of subgenotype Z.1 viruses could be distinguished phylogenetically from the NS segments of subgenotype Z.2 and Z.3 viruses, which could not be distinguished from each other. Some mutations that distinguished the NS genes resulted in amino acid replacements.

In contrast to the multiple HA clades of genotype Z viruses, the three subgenotypes are characterized by distinct HA clades. All except one of the 63 subgenotype Z.2 viruses had an HA from clade 1; the 175 subgenotype Z.1 viruses possessed an HA gene of clade 2.2(0.1), with seven exceptions; and the 31 subgenotype Z.3 viruses had HA genes of clades 2.1.2 and 2.1.3 with two exceptions. Fig. 4 provides a summary of our model for genotype evolution relating the vast majority of the avian H5N1 viruses in our study.

Subgenotype Z.1 and Z.2 viruses persisted and expanded. By contrast, subgenotype 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 the genotypes descending from Z.3 have been previously identified. (Supplementary Material lists viruses for which genotypes were assigned here and by others.) Clade 2.3.4 (Fujian-like) viruses were originally described as genotype Z (Smith et al., 2006a), but later reclassified as genotype V (Duan et al., 2008). Recently, clade 2.3.4 HA was also observed among viruses of genotype G (Smith et al., 2009). Our analysis provides a more detailed picture of the evolution of clade 2.3.4 viruses, which form the majority of genotypes 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 (Chen et al., 2006b), a reassortment event that we also detect (see subgenotype Z.3 to genotype 33, Fig. 3). In addition, we detect a reassortment event in the PB2 gene, resulting in genotype 35 that 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. (2008) reported that genotypes Z and G differ in the PB2 segment. 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.

We noted three reassortant viruses, each with a distinct genotype, which could be traced to reassortment of a single internal segment of Z.1 viruses (not shown in Fig. 3 for reasons of clarity); each virus had clade 2.2 HA. However, none of these three novel genotypes persisted. We could trace no reassortants to subgenotype Z.2. Sixteen additional viruses (not shown) either had unclassifiable segments because of low bootstrap support, or differed by more than three segments from any virus (Fig. 3) and hence could not be assigned to the evolutionary pathway; some of these unclassifiable segments appeared to include novel genetic material.

Molecular changes in internal genes

Our analysis revealed a number of potentially significant molecular changes in the PB2, PA and NS1 proteins that appear to be associated with specific genotypes or subgenotypes. We describe some of these here.

The viruses of subgenotype Z.1 (including Qinghai-Lake viruses and their descendants) possess a number of ‘unconventional’ molecular changes. As described by others, Z.1 viruses possess the ‘human-like’ Lys at PB2-627 that confers increased pathogenicity in some mammalian species (Hatta et al., 2001; Chen et al., 2005; Liu et al., 2005). In addition, we detected other unconventional changes in the PB2 protein of Z.1 viruses, some of which (Iso-to-Thr at position 147, Arg-to-Gln at position 368 and Asp-to-Asn at position 390) were also found in
the PB2 protein of genotype 35 viruses (Fig. 3). Since both groups of viruses have circulated and geographically expanded since 2005, one may hypothesize that the unique amino acids in their PB2 proteins are of biological significance, perhaps by increasing the replicative ability of these two groups of viruses. However, experimental...
Fig. 3. Model of the evolution of avian H5N1 influenza viruses from 2005 to 2008. The details of bars, line styles, colours, NS deletion and descriptions of genotypes are as described in Figs 1 and 2. Genotype Z viruses have now differentiated into three distinguishable subgenotypes, 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 geographical origin of the respective (sub)genotypes.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Representative virus</th>
<th>History of generation</th>
<th>Viruses per genotype</th>
<th>Clade</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Z.1)</td>
<td>A/peregrine falcon/HK/D0028/04</td>
<td>Evolved from Z by mutations in PB2, PB1, PA, NS</td>
<td>175</td>
<td>2.2</td>
</tr>
<tr>
<td>(Z.2)</td>
<td>A/ck/Yunnan/125/03</td>
<td>Evolved from Z by mutations in PB2, PB1, PA, NS</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>(Z.3)</td>
<td>A/ck/Hunan/782/03</td>
<td>Evolved from Z by mutations in PB2, PB1, PA, NS</td>
<td>31</td>
<td>2.1.3</td>
</tr>
<tr>
<td>(31)</td>
<td>A/Guangxi/3055/05</td>
<td>A/ck/Guangxi/1311/04 (H5N1)-like NS</td>
<td>7</td>
<td>2.3.3</td>
</tr>
<tr>
<td></td>
<td>Unknown PA (likely of genotype Z origin)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(32)</td>
<td>A/comon magpie/HK/2125/06</td>
<td>A/ck/Guangxi/1311/04 (H5N1)-like PB2</td>
<td>9</td>
<td>2.3.4</td>
</tr>
<tr>
<td>(33)</td>
<td>A/ck/Fujian/1094/05</td>
<td>A/ck/Guangdong/74/04 (H5N1)-like PA</td>
<td>30</td>
<td>2.3.4</td>
</tr>
<tr>
<td>(34)</td>
<td>A/ck/Guangxi/3154/05</td>
<td>A/ck/Guangxi/1311/04 (H5N1)-like PB2</td>
<td>38</td>
<td>2.3.2</td>
</tr>
<tr>
<td>(35)</td>
<td>A/ck/Guangxi/5165/05</td>
<td>Sub-lineage Z.1-like PB2</td>
<td>52</td>
<td>2.3.4</td>
</tr>
<tr>
<td>(36)</td>
<td>A/ck/Fujian/688/06</td>
<td>A/ck/Guangdong/74/04 (H5N1)-like PB2</td>
<td>4</td>
<td>2.3.4</td>
</tr>
<tr>
<td>(37)</td>
<td>A/ck/Yunnan/5133/05</td>
<td>A/blackbird/Hunan/1/04 (H5N1)-like PA</td>
<td>6</td>
<td>2.2.7</td>
</tr>
</tbody>
</table>
data that test this hypothesis are not available at this point.

Unusual molecular changes in both NS1 and PB1-F2 proteins are associated with genotype 31. This group of viruses possesses an NS gene encoding a premature stop codon that abrogates expression of the two C-terminal amino acids of NS1. The four C-terminal amino acids of NS1 form the PDZ-domain ligand motif (Obenauer et al., 2006) that contributes to virulence (Jackson et al., 2008). The truncation of the PDZ-domain ligand 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 aa. PB1-F2 functions as a pro-apoptotic factor (Chen et al., 2001; Zamarin et al., 2005), enhances inflammation in mice and increases the severity of secondary bacterial infections (McAuley et al., 2007). Moreover, it interacts with the PB1 protein to retain it in the nucleus and may, through this mechanism, affect virulence (Mazur et al., 2008). Currently, it is not clear if the C-terminal truncation of NS1 and the truncation of PB1-F2 are functionally linked.

Collectively, these examples demonstrate the power of high-resolution genotyping in identifying evolutionarily significant mutations, and even potentially co-evolving mutations.

**DISCUSSION**

We present a detailed and comprehensive analysis of the evolution of highly pathogenic avian H5N1 influenza viruses, which allowed us to propose a model for reassortment of internal genes involved in the emergence of the currently dominant groups of viruses, i.e. sub-genotype Z.1 viruses (HA clade 2.2), subgenotype Z.2 viruses (HA clade 1) and genotype 35 viruses (HA clade 2.3.4). Further, our genotyping highlighted molecular changes in the polymerase and NS1 proteins, and a possible role for interactions between HA and internal proteins, in the emergence of these dominant variants.

Our studies echoed earlier fundamental observations about the role of reassortment in the evolution of H5N1 influenza viruses (Guan et al., 2002a; Li et al., 2004; Chen et al., 2004, 2006b; Smith et al., 2006a; Duan et al., 2008). 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 (Xu et al., 1999), is characterized by an allele B NS segment, while almost all of the current H5N1 viruses possess NS segments of allele A. We observed that, after their appearance in 2001, genotype Z viruses underwent frequent reassortment events, predominantly with H5N1 viruses. A few reassortment events involved genes previously undetected in H5N1 viruses, while present in non-H5N1 viruses, which is congruent with observations that avian influenza viruses of other subtypes likely co-circulated with H5N1 viruses in South-east Asia at that time (Cheung et al., 2007; Xu et al., 2007). Notably, these reassortment events produced only transient genotypes, suggesting that genotype Z was highly competitive.

In addition, our analyses provided new insight into the role of reassortment in the evolution of avian H5N1 viruses. Specifically, our model pointed to a distinguished role for
reassortment of the PB2, PA and NP genes. This contrasted with earlier studies of mixed subtypes of avian influenza virus that suggested that all segments may participate equally in reassortment (Macken et al., 2006; Dugan et al., 2008). The polymerase complex (PB2, PB1 and PA proteins) is now recognized as a determinant of pathogenicity (Hatta et al., 2001; Li et al., 2005; Gabriel et al., 2005; Salomon et al., 2006; Van Hoeven et al., 2009; Watanabe et al., 2009), although neither the contributions of the individual polymerase proteins nor the underlying mechanisms are fully understood. Replacement of the polymerase genes in the early phase of avian H5N1 virus evolution may have affected the replication levels of the respective viruses and hence their ability to compete with co-circulating viruses of other genotypes or subtypes (Crescenzo-Chaigne et al., 1999; Naffakh et al., 2000; Massin et al., 2001; Shiniya et al., 2004).

In our analyses, we observed that the persistence of genotype Z viruses resulted in the accumulation of point mutations, leading to three distinguishable subgenotypes (designated here as Z.1, Z.2 and Z.3). These subgenotypes 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 (Smith et al., 2006b). The divergence into subgenotypes was apparent for the polymerase and NS genes, but not for the NP and M genes. This finding suggested that the accumulated point mutations are not merely the result of the error rate of the viral replication complex, but, at least in part, reflected evolutionary pressure.

Interestingly, our analysis suggested that changes in the make-up of the internal genes often coincided with changes in the HA clade. (It is important to recall that HA was not used in determining genotypes.) In the 9 years of evolution covered by our study, three qualitatively different patterns of association between HA clade and genotype stood out: (i) persistent HA, but multiple genotypes on the A/duck/Anyang/AVL-1/01 pathway (genotypes 3–6 and 25–30). This pathway included 10 genotypes with 15 reassorting segments. Viruses on this pathway had a closely related HA from clade 0. (ii) Vigorous HA evolution in genotype Z (genotypes 14 and Z). The progenitor of genotype Z (A/duck/Guangxi/50/2001) had a clade 9 HA. Seven new HA clades were observed among genotype Z viruses, the only example in our analysis of many HA clades associating with a single genotype. (iii) Persistent HA – persistent subgenotypes (Z.1 and Z.2). Rare reassortants of subgenotypes Z.1 and Z.2 were observed and HA changed only from 2.2 to 2.2.1 in Z.1 viruses. A plausible explanation for these substantially different behaviours is that HA and internal proteins interact to a greater extent than previously recognized. Frequent changes in HA clade or internal genes early in H5N1 virus evolution may reflect the need to find a good ‘match’, leading to recent H5N1 viruses characterized by greater genetic stability. Another observation suggesting co-ordination between HA and internal genes is the recent expansion of HA clade 2.3.4 in conjunction with genotype 35, but not in conjunction with the other three genotypes (32, 33 and 36) also carrying this HA clade (Fig. 3). It is possible that HA and the internal genes are finely tuned to each other.

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. Researchers have often resorted to the comparison of pairs of viruses that differ in their virulence or pathogenicity. Paired comparisons can identify 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 (Sweet et al., 2004). Recently, large-scale sequencing approaches have been employed to identify signature amino acid changes in viral genomes. Such an approach identified the PDZ-domain ligand motif in the NS1 protein (Obenauer et al., 2006); however, sequence alignments do not take into account the genetic relationships among viruses and hence are not well suited to identify changes that correlate with significant evolutionary events such as reassortment. Hence, changes (such as the truncated NS1 and PB1-F2 proteins in genotype 31 viruses) may appear random when viewed in a simple summary (such as in sequence alignment) of a large collection of viruses. By contrast, our analysis suggests that these changes may reflect selective pressure in conjunction with a reassortment event (from subgenotype Z.3 to genotype 31; see Fig. 3).

In summary, our studies led us to a rich model of avian H5N1 virus evolution, involving reassortment, significant gene diversification by point mutations and a role for the interaction of HA and internal genes in the emergence of dominant variants. Such knowledge may help in predicting future events in avian H5N1 virus evolution.

**METHODS**

**Data.** The study was limited to avian influenza A viruses with >90% of full-length sequences for each internal gene segment and HA. The study genotyped highly pathogenic avian influenza A (H5N1) viruses from 2000 to 2008 (‘test viruses’). Genetic data were drawn from GenBank (Benson et al., 2008) with the most recent update of the analyses occurring in May 2009.

**Alignments and phylogenetic trees.** Alignments of nucleotide sequences were generated using the profile alignment option of CLUSTAL W, 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 neighbour-joining trees, in order to achieve high-throughput (in our earlier study, Macken et al., 2006), 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 (Felsenstein, 2004); these were fitted by maximizing likelihood using PhyML (Guindon & Gascuel, 2003).
Two-time test. The two-time test for genotyping influenza viruses was described in detail in our earlier publication (Macken et al., 2006). Briefly, ‘test’ viruses in one time period are genotyped relative to ‘reference’ viruses in an earlier, neighbouring time period. Reference viruses represent lineages from which genetic material is drawn in the reassortment process. ‘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. 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.

The two-time 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 Supplementary Material.

Genotyping. We genotyped the PB2, PB1, PA, NP, M and NS segments of all test viruses using the two-time test. We identified the clade of the HA segment of each H5N1 virus according to the unified nomenclature system (WHO, 2008). Because some of our ‘test’ viruses were not included in this analysis (WHO, 2008), we inferred the maximum-likelihood phylogeny of HA of all test viruses under the HKY model of evolution with maximum-likelihood optimization using PhyML (Guindon & Gascuel, 2003). If an unclassified HA had strong support for associations with a classified HA, we assign the appropriate HA clade to the former HA. Otherwise, HA was not classified.

ACKNOWLEDGEMENTS

C. A. M. and M. A. G. gratefully acknowledge the support of the US Department of Energy through the LANL/LDRD Program for this work.

REFERENCES


(H5N1) virus: similarity of its hemagglutinin gene to those of H5N1 viruses from the 1997 outbreaks in Hong Kong. Virology 261, 15–19.
