

Published online ahead of print on 10 July 2006 as DOI 10.1099/vir.0.81454-0

**Genotype turnover by reassortment of replication complex genes from avian
*Influenza A virus***

Catherine A. Macken,¹ Richard J. Webby² and William J. Bruno¹

¹Theoretical Biology and Biophysics, Los Alamos National Laboratory, T-10 MS-K710,
Los Alamos, NM 87545, USA

²Department of Infectious Diseases, St Jude Children's Research Hospital, 332 N. Lauderdale,
Memphis, TN 38105-2794, USA

Correspondence

Catherine A. Macken
cmacken@lanl.gov

Supplementary material is available.

Reassortment among the RNA segments of *Influenza A virus* caused the two most recent human influenza pandemics; recently, reassortment has generated viral genotypes associated with outbreaks of avian H5N1 influenza in Asia and Europe. A statistical analysis has been developed for the systematic identification and characterization of reassortant viruses. The analysis was applied to the genes of the replication complex of 152 avian influenza A viruses isolated between 1966 and 2004 from predominantly terrestrial and domestic aquatic avian species. The results indicated that reassortment among these genes was pervasive throughout this period and throughout both the Eurasian and North American lineages of the virus. Evidence is presented that the circulating genotypes of the replication complex are being replaced continually by novel genotypes created by reassortment. No constraints for coordinated reassortment among genes of the replication complex were evident; rather, reassortment almost always proceeded one segment at a time. A maximum-likelihood estimate of the rate of reassortment was derived. For significantly diverged Asian avian influenza A viruses from the period 1991–2004, it was estimated that the median duration between creation of a new genotype and its next segment reassortment was 3 years. Reassortments that introduced previously unobserved influenza genetic material were detected. These findings point to substantial potential for rapid generation of novel avian influenza A viruses, emphasizing the importance of intensive surveillance of these host species in preparation for a possible pandemic.

INTRODUCTION

Influenza A virus is an enveloped, negative-sense RNA virus with a genome in eight segments. Segmentation of the viral genome permits evolution by a process called reassortment, an outcome of co-infection of a permissive host cell by at least two different influenza A viruses. The resulting novel, 'reassortant' virus is composed of a mixture of RNA segments from the parent genomes.

Reassortment can lead to dramatic changes in the viral phenotype, producing, for example, the pandemic strains of 1957 and 1968. All genes of the influenza A viral genome can reassort and the fitness of internal gene reassortants varies (Murphy *et al.*, 1984; Snyder *et al.*, 1987; Clements *et al.*, 1992). Since 1997, avian-adapted H5N1 viruses, which are reassortants of avian influenza viral lineages (Guan *et al.*, 1999; Hiromoto *et al.*, 2000; Hoffmann *et al.*, 2000; Li *et al.*, 2004), have infected humans, causing mortality in many cases. To date, these viruses are at best poorly transmissible from human to human and a pandemic strain has not yet emerged. Of eminent concern is the possibility that avian H5N1 viruses may evolve to become capable of reassorting successfully with co-circulating human influenza A viruses, leading to a highly pathogenic, human-transmissible reassortant.

In general, few details are known about capacity for reassortment (Scholtissek *et al.*, 2002). Whilst it is well accepted that co-ordination of the evolution of the haemagglutinin and neuraminidase gene segments is needed for optimal viral growth (Wagner *et al.*, 2002), corresponding information on other influenza virus gene segments is scant.

Reassortment is identified phylogenetically, when segments of a viral genome have inconsistent associations with distinct clades of viruses in segment-specific phylogenies. All genes of avian influenza A viruses divide into two broad phylogenetic lineages, labelled North American and Eurasian after the two major flight paths of wild aquatic birds. Viruses isolated from domestic or wild avian hosts in a particular geographical region, such as South-East Asia, usually segregate phylogenetically with other viruses collected from other geographical regions through which the same flight path passes. This suggests that wild aquatic birds were the historical, and perhaps ongoing, source of influenza genetic material in avian livestock. The two flight paths are almost completely geographically separated; reassortment between their viral lineages is rare (Wallensten *et al.*, 2005) and readily recognizable.

Less-diverged lineages are distinguishable phylogenetically within the North American and Eurasian lineages. The observed reassortment in avian influenza A H5N1 viruses occurs among these subtler lineages. Whilst it may be possible to unravel influenza reassortment within hosts with few viral lineages, analysing reassortment in avian hosts with multiple co-circulating viral lineages is substantially more complex. The extreme lack of congruence of the phylogenies in Fig. 1 of avian influenza PB2, PB1, PA and NP genes, the four genes of the virus replication complex, suggests this difficulty.

We developed a novel statistical methodology for examining the details of reassortment. Our fundamental insight was that, whilst analysis of reassortment using phylogenies from a single time period may be too complex, by considering two consecutive periods, we could use

the first (T1) to establish a baseline of circulating genotypes, then characterize the reassortant status of viruses from the second (T2) relative to the potential donor viruses in T1. Using our methodology, we studied reassortment among the PB2, PB1, PA and NP genes in avian hosts. The large variety of avian host species (Alexander, 2000) provides many opportunities for detectable reassortment. The compact structure of the replication complex (Area *et al.*, 2004) might suggest restrictions on the reassortment of particular combinations of these genes. As PB2 has been associated with virulence (Hatta *et al.*, 2001), the replication complex is an important context for our analysis of reassortment. Our approach proved to provide a powerful means to investigate the role of evolution by reassortment in the ecology of *Influenza A virus*. By understanding the factors that underlie reassortment, it may be possible to identify hosts in which it occurs more frequently and conditions that favour genetic exchange, potentially identifying areas where changes in the management of domestic poultry populations can reduce the genetic diversity of circulating viruses.

METHODS AND RESULTS

Our study had two phases. First, we carried out our two-time test [(i) below] to classify the reassortant status of the replication complex of viruses in time period T2 relative to the genotypes (i.e. observed combinations of lineages for the four segments) circulating in the preceding time period, T1. We also observed which, if any, segments reassorted in a coordinated fashion. Second, by using the results of (i) to formulate a model for the reassortment process, we estimated the rate of reassortment among the segments of the replication complex [(ii) below].

(i) Two-time test

The two-time test was performed in two steps: first, categorization of the genotypic composition of the circulating viral population in T1 by constructing reference trees; second, ascertainment of the reassortment status of T2 viruses by measuring bootstrap support for their placement on the segment-specific reference trees. A T2 virus was declared a relative reassortant if, when added to segment-specific reference trees, it associated with >70 % bootstrap support with different viruses in at least two of the trees (see note below). Otherwise, the T2 virus was declared a relative non-reassortant. The modifier 'relative' makes explicit that the reassortant status of a T2 virus is evaluated relative to the T1 viral genotypes. Our two-time test construction is analogous to phylogeny-based tests for recombination in non-segmented genomes (Dorman *et al.*, 2002; Chare *et al.*, 2003).

Note: we have assumed that the dataset does not include recombinant segments: recombination is apparently rare in influenza viruses. If recombinants are present in our reference trees, our results still hold except that, in some instances, declared reassortants are probably recombinants.

We describe here our two-time test of more recent, post-1990 viruses. We also analysed older, pre-1995 viruses (see Supplementary Material). The selection of T1 and T2 is at our discretion. We noted that, in domestic avian viruses, lineages of PB2, PB1, PA or NP rarely persisted for 10 years, suggesting an upper limit on the length of time that lineages might co-circulate (a prerequisite for reassortment). Hence, we chose T1 and T2 to be approximately 10 years.

Data selection. Sequences were selected from those available in the Influenza Sequence Database (ISD) (Macken *et al.*, 2001) in July 2005. We selected only viruses for which all four segments were fully or mostly sequenced. Stored alignments from the ISD were used: no editing was necessary. Supplementary Table S1 lists our final selection of 152 viruses (16 pre-1995 and 145 post-1990 viruses).

Choice of taxa for reference trees. Reference trees were required (i) to include as many as possible of the distinguishable genotypes of T1 viruses, in order to be able to classify accurately 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 could be unequivocal, and (iii) to have an outgroup unrelated by reassortment to any of the T1 or T2 viruses.

Choice of outgroup. Natural choices for an outgroup, such as A/FPV/Rostock/34, did not meet the criterion of being unrelated by reassortment to any of the T1 viruses. As very distant outgroups (e.g. *Influenza B virus*) lead to reduced accuracy of inference of topologies among ingroup viruses (Korber *et al.*, 2000), we inferred ancestral sequences of PB2, PB1, PA and NP from phylogenies of approximately 100 older influenza A viruses from swine, avian, human and equine hosts (by using PAML; Yang, 1997). Inclusion of viruses from non-avian hosts led to an ancestor more distant than the common ancestor of avian influenza viruses, but closer than influenza B viruses. These ancestral sequences met our criterion for outgroup.

Inference of T1 (1991–2000) reference trees. Seventy-two viruses had sufficient data to be considered for reference trees. We inferred their phylogenies (see Fig. 1) by using PAUP* (Swofford, 1993) with the minimum-evolution criterion under the HKY85 model of evolution (Hasegawa *et al.*, 1985). To focus our analysis on reassortment among distinguishable genotypes, we selected one representative of each as follows. Starting from the tips of the phylogenies of Fig. 1, we identified clusters of viruses collocated in all four segment-specific phylogenetic trees with high bootstrap support. We selected a single member of the cluster to represent the cluster genotype. For example, A/goose/Guangdong/1/96, A/goose/Guangdong/3/97, A/goose/Hong Kong/ww26/2000, A/goose/Hong Kong/ww28/2000, A/duck/Guangdong/07/2000 and A/duck/Guangdong/12/2000 clustered consistently with high bootstrap support in the four segment-specific trees. Others of the 72 T1 viruses clustered with these six viruses in some, but not all, of these trees; for example, A/duck/Zhejiang/11/2000

joined the cluster in only PB1, PA and NP trees. We chose A/goose/Guangdong/3/97 to represent the genotype of these six consistently clustered viruses. This grouping and selecting reduced the T1 dataset to 28 viruses, each representing a distinguishable genotype (Fig. 1). The phylogenies of these 28 representative viruses became 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. However, 16 of the 28 viruses in the draft reference trees attached at deeper nodes with low (<70 %) bootstrap support in one or more of the segment-specific draft reference trees; these viruses (highlighted in blue in Fig. 1) were removed to produce our final reference trees (Fig. 2), but were kept as candidate donor segments.

We denoted the collection of 28 (highlighted in yellow or blue in Fig. 1) as the reference set, a subset of which contained the 12 viruses (yellow in Fig. 1) of the reference trees. When we later characterized the genotypes of T2 viruses, we first looked for donor segments among the viruses of the reference tree. If no 'yellow' donor was apparent (i.e. if the attachment point was basal to multiple reference-tree viruses), we searched the 'blue' viruses for the donor. To make searching 'blue' viruses for donors easier, we determined the placement of the 'blue' viruses individually on the reference trees. All trees with 12 'yellow' and a single 'blue' virus had >70 % bootstrap support for the placement of the 'blue' virus (see Supplementary Table S3). All 'blue' viruses have genotypes that differed from each other and from those of the 'yellow' viruses.

Inclusion of the 'blue' viruses thus reduced the chances of mislabelling a T2 virus as a relative reassortant simply because its genotype, whilst present in the T1 population, was not in our reference trees. Our conclusion that the T2 virus underwent reassortment between times T1 and T2 would then be incorrect; the reassortment event would have been earlier. It is implausible, however, that our results were influenced substantially by missing a large number of very old reassortants in T1 surveillance; after much more than 10 years, lineages are not recognizable and could not be placed on a reference tree with high bootstrap support.

All except one node (deep in the PB1 reference tree) of the segment-specific reference trees had a bootstrap value >70 %. We elected to retain the viruses that were descended from this single node in order to improve the precision of placement of T2 viruses on the reference trees. We believe that our conclusions are robust to low bootstrap values on deep nodes, because deep nodes represent historical, not contemporary, evolutionary events: for our purposes, we are concerned with reassortment among contemporary viruses.

We checked the sensitivity of our reference trees to analytical method by reestimating with PAUP using different models of evolution and methods of optimization, and also with WEIGHBOR (Bruno *et al.*, 2000). The topologies of the reference trees for all methods were the same and bootstrap support was similar for every node; the assessment of reassortment described below was the same for all methods.

Construction of the reference trees for T1 viruses did not demand that they contained non-reassortant viruses. The consensus of the four segment-specific reference trees (calculated by using PHYLIP; Felsenstein, 1993) showed disparate phylogenies (Fig. 2). We concluded that

reassortment was involved in the evolutionary relationships among viruses of the 1991–2000 reference trees.

Classification of T2 (2001–2004) viruses relative to T1 reference trees. The ISD contained 74 viruses with sufficient data to be analysable for reassortant status. In practice, those subsets of the 74 that clustered closely with high bootstrap support in all four phylogenies of 2001–2004 viruses would have the same reassortant status. Thus, we analysed the reassortant status of one representative of each cluster, resulting in analysis of 31 T2 viruses.

In assessing relative reassortant status, viruses either associated with a single reference virus (always with >70 % bootstrap support and almost always with >90 % bootstrap support) or attached with >70 % bootstrap support basal to a clade of viruses in the reference tree, sometimes with disruption in the subtree at the point of attachment (e.g. see Fig. 3). Thus, even in the presence of basal attachments, classification of relative reassortant status of T2 viruses was well supported and is summarized in Table 1. Of the 31 T2 viruses tested, *A/duck/Shanghai/08/2001* was a sister to *A/goose/Guangdong/3/97* in all segment-specific reference trees and was declared a relative non-reassortant; the other 30 were declared relative reassortants, representing 17 distinct Eurasian genotypes and four distinct North American genotypes. Note that, whilst *A/duck/Shanghai/08/2001* is a non-reassortant relative to T1 viruses, the reassortant composition of *A/goose/Guangdong/3/97* (as evident from its discordant placement among segment-specific reference trees) emphasizes that non-reassortant classification of *A/duck/Shanghai/08/2001* is relative, not absolute. Relative to viruses earlier than T1, *A/duck/Shanghai/08/2001* was a reassortant.

Basal attachment of a T2 virus might indicate missing genetic data due to, say, the sparse database on the American lineage. In contrast, genotypes of Eurasian viruses were represented quite densely in the reference set. Most basal attachments of T2 Eurasian viruses matched the basal attachment of a 'blue' virus when it was earlier characterized (see Supplementary Table S3). For example, the PA of *A/chicken/Hong Kong/FY150/2001*, classified as (CkKor,GsGu,Ostr) in Table 1, matched the classification of the PA of *A/duck/Zhejiang/52/2000*. Hence, we constructed a new PA reference tree with both *A/chicken/Hong Kong/FY150/2001* and *A/duck/Zhejiang/52/2000* added (Fig. 3c). In every case, the T2 virus had at least 70 % bootstrap support for being a sister to the 'blue' virus.

Two types of each of PB2 and PA segments with basal attachments (see bold type in Table 1) did not associate closely with any 'blue' virus. To examine the extent to which these segments might be novel, we compared the PB2 and PA genes of *A/duck/Shanghai/35/2002* with the database of human-, avian- or swine-adapted influenza A viruses having at least 2000 nt of PB2 or PA sequence. The PB2 of *A/duck/Shanghai/35/2002* differs from the four viruses to which it was basal by between 134 and 197 nt and had two unique positions, 338A and 443E. Positions 251K and 588T were shared with only a cluster of predominantly European (251K) or North American (588T) swine-adapted viruses. The PA of *A/duck/Shanghai/35/2002* differs from the five viruses to which it was basal by between 165 and 207 nt and had two unique positions,

623G and 679R. Therefore, the basal attachments of these Eurasian PB2 and PA genes coincided with novel sequence details rather than with deficiencies in the reference set.

Among the tested T2 viruses, we observed a genotype having PB2 from one T1 genotype and the remaining three segments from a different T1 genotype (Table 1). Viruses that had two segments from one T1 genotype had the other two segments from two different T1 genotypes. We also observed T2 genotypes composed of four distinct T1 genotypes. These observations point to reassortment of these genes occurring in steps, one segment at a time and independently of other genes of the complex.

The detailed characterizations of the replication complex genotypes in Table 1 allowed inference of the likely reassortment events that led to their creation from the T1 genotypes. A most parsimonious phylogeny of these 2001–2004 genotypes (assuming independent, single-segment reassortment, as shown above) is given in Fig. 4; two other equally parsimonious phylogenies involve arrangement of the trio of viruses that descend from *A/ostrich/South Africa/9508103/95*. In Fig. 4, only one representative is shown for each genotype in each year that it was observed. The lack of persistence of genotypes from T1 and short periods of persistence of genotypes in T2 point to rapid replacement of old replication complex genotypes by newly created replication complex genotypes among these viruses.

(ii) Inference of the rate of reassortment

Our estimate of the rate of reassortment was based on data on the years at which reassortant genotypes were observed, and the estimated year of creation of the parental genotype. The Supplementary Material contains our likelihood function from which we obtained the maximum-likelihood estimate (MLE) of the rate of reassortment. Our estimation procedure depended on three components, explained below: the selection of data containing information on the rate of reassortment, a model for evolution by reassortment and a probability distribution for the variables in our likelihood function.

Selection of data. We required a dataset with sufficiently dense coverage of T1 and T2 replication complex genotypes to be able to track accurately the emergence by reassortment of T2 genotypes from T1 genotypes. Adequate data were available from post-1990 viruses isolated from Asian terrestrial avian host species. We used a single representative of a genotype for each year that the genotype was observed to track the emergence and persistence of genotypes (Fig. 4). With the exception of two viruses, whose origin could not be ascertained, the replication complex genotype of all T2 Asian viruses was descended by reassortment from either the '*A/goose/Guangdong/3/97*' genotype or the '*A/ostrich/South Africa/9508103/95*' genotype. We used these data to obtain (a) the number of reassortment events that changed a T1 genotype into a T2 genotype and (b) the time taken for these reassortment events to occur.

(a) Inference of the number of reassortment events is straightforward from Fig. 4. For example, the genotype of *A/chicken/Hong Kong/YU822.2/01* differs from the genotype of *A/goose/Guangdong/3/97* in one segment (PB2). Additional reassortment of PB1, PA or NP in

the A/chicken/Hong Kong/YU822.2/01 genotype led to three new genotypes (one in each of 2001, 2002 and 2003), differing from the A/goose/Guangdong/3/97 genotype by two segments.

(b) The number of reassortment events counted in (a) took place between the time that the T1 genotype emerged and the date at which the T2 virus with the reassorted genotype was observed. Genotypes in T1 emerged possibly years before they were observed, but necessarily after the most recent common ancestor (MRCA) of all lineages from which the genotype is composed. This observation allowed us to estimate a bound on the date of emergence, which we denoted C . As we saw in (i) above, the T1 viruses in our study are reassortants. Hence, we estimated C by the date of the youngest MRCA of the PB2, PB1, PA and NP lineages involved in the T1 genotype.

To estimate the date of the MRCA of viruses in a lineage, we applied TipDate (Rambaut, 2000) to all full-length PB2, PB1, PA and NP sequences from Eurasian non-aquatic avian viruses available in July 2005 in the ISD (each dataset included 120–180 sequences), with A/FPV/Rostock/34 as the root [TipDate uses the dates of tips on a viral sequence phylogeny to estimate the annual rate of nucleotide change of these sequences under the assumption of a molecular clock (Wilson *et al.*, 1977)]. To reduce the influence of selection due to host change or reassortment on the estimated annual rate of nucleotide change, we used third-codon positions only.

TipDate estimated 3.69×10^{-3} , 3.80×10^{-3} , 4.52×10^{-3} and 3.85×10^{-3} changes $\text{nt}^{-1} \text{ year}^{-1}$ for PB2, PB1, PA and NP, respectively. In each of the four trees to which we had applied TipDate, we used the estimated rates of change to back-estimate the date of the MRCA for the subtree containing all T1 viruses from the 'A/goose/Guangdong/3/97' genotype, together with all T2 viruses whose PB2/PB1/PA/NP associated with A/goose/Guangdong/3/97 in T1 reference trees. We obtained 1991, 1980, 1978 and 1983, respectively, giving an estimate of $C=1991$ for the genotype represented by A/goose/Guangdong/3/97. Similarly, we estimated the date of the MRCA of TipDate subtrees containing all 'A/ostrich/South Africa/9508103/95' genotype T1 viruses and all T2 viruses whose PB2 or PB1 associated with A/ostrich/South Africa/9508103/95 in T1 reference trees. We obtained 1986 and 1987, respectively. We did not estimate the MRCA for the PA or NP of A/ostrich/South Africa/9508103/95; the subtrees for PA and NP contained T1 viruses only. As we wanted all estimates of MRCA to be based on similar amounts of clonal variation, which affects the precision of the estimate of the MRCA, we estimated the MRCA only for subtrees containing both T1 and T2 viruses. Hence, $C \geq 1987$ for the genotype represented by A/ostrich/South Africa/9508103/95. We used $C=1987$, leading to a possibly conservative (low) estimate of the rate of reassortment for the descendants from this genotype.

Model of evolution by reassortment. We assumed that segments of the replication complex reassort independently, one segment at a time, which had strong support from (i) above, occurring at a constant rate. Specifically, we assumed that each segment of this complex reassorts independently of the other three segments with constant annual probability $(1-\rho)$; we wish to estimate ρ .

Probability distribution over the number and timing of reassortment events. Under the model described above, when a novel genotype is created in year C , the probability that n out of its four segments have reassorted by time $C+R$ is Binomial($4, \rho^R$), that is, proportional to $\rho^{(4-n)R}(1-\rho^R)^n$ (n segments have, and $4-n$ have not, reassorted within the R years). We ascertain R and n for each virus in Fig. 4; the likelihood function of the complete dataset is built from these binomial probabilities (see details in Supplementary Material). We maximized the likelihood function with respect to ρ to obtain a conservative MLE, $\hat{\rho}$. Our conservative estimate of ρ was 0.9473 (a computer program for this calculation is available upon request).

We denote by T_R the time in years between creation of a new genotype at C and the first reassortment with a biologically distinct lineage to create a new, viable (i.e. transmissible) genotype. Then, the probability that $T_R=i$ years is $\rho^{4(i-1)}(1-\rho^4)$ (the probability that no reassortment in a genotype occurs for $i-1$ years, followed by at least one reassortment event in the i^{th} year) for $i=1, 2, 3, \dots$

Using our estimate of ρ , the median value of T_R is calculated to be 3 years and the chance that T_R exceeds 7 years is 22 %. That is, when a new genotype is created, 50 % of the time a segment will have reassorted within 3 years of the creation, although there is a long upper tail on this distribution. Uncertainty in the estimate of T_R arises from uncertainty in the estimates of segment-specific rates of evolution, which are calculated by TipDate. However, the uncertainty in the estimated rates of evolution affected the estimates of the MRCAs by less than 1 year, which had a negligible effect on the estimate of the median value of T_R .

DISCUSSION

Outstanding fundamental questions exist about restrictions on, and rate of, reassortment and the role of reassortment in the composition of the avian influenza viral population. Such information is prerequisite to understanding how viral lineages emerge successfully and in aiding prediction of virus evolution. We developed a ‘two-time test’ that enabled us to analyse the evolution of genotypes by reassortment. Using our two-time test, we showed that, throughout a 30 year history, the segments of the replication complex of avian influenza viruses from both the North American and Eurasian lineages appear to have reassorted one segment at a time with no obvious constraints on their ability to reassort successfully among significantly diverged sublineages. This flexibility of reassortment echoed recent observations (Hatchette *et al.*, 2004) that all internal segments of *Influenza A virus* in feral Canadian ducks reassort without apparent limitation.

Our analysis indicated that reassortment led to replacement of older genotypes by newly emerging genotypes. Of the 28 viruses (representing 84 in the database) whose reassortant status was tested, only two were non-reassortants relative to earlier genotypes, 33 were relative reassortants and two could not be classified. Thus, these data supported the scenario of Fig. 5(b), with virtually complete turnover by reassortment over a 10 year period. If the alternative scenario of clonal evolution with occasional, short-lived reassortants (Fig. 5a) were to hold, we would expect our testing to reveal a substantial majority of non-reassortant viruses. Whilst

influenza sequence sampling is notoriously non-random, we find the same trend in our results over datasets of older and newer viruses from each of the two major avian viral lineages. This robustness suggests that increased sampling is unlikely to contradict the basic conclusions of our two-time test.

The impetus for our new methodology is the recognition that a set of segment-specific trees for a single selection of viruses can distinguish reassortants from non-reassortants with certainty only when the incidence of reassortment is low. Our two-time test ensured that reassortment in one period was assessed accurately relative to the genotypes of an earlier period.

Our approach bears similarities to the typical description of genotypes as mixtures of viral lineages (e.g. the 'goose/Guangdong' lineage) that appear to be prevalent in an earlier time period. However, the subjective nature of grouping viruses either genetically (as typical) or (more precisely) on the basis of protein sequences (e.g. Obenauer *et al.*, 2006) generally recognizes only substantially diverged groupings, giving a rather low-resolution view of reassortment. We avoided subjective clustering by deriving reference trees with strong statistical support at the level of resolution of significant evolutionary events, i.e. reassortment.

A reference tree could contain a single representative virus from of each of the North American and Eurasian flyways and be used to detect relative reassortants between these major lineages. The detail of the reference tree determines the precision of identification of relative reassortants. Uniqueness is not necessary for testing candidate reassortants. The attribute of reference trees that makes them effective is their well-supported topologies. We found infrequent changes in our reference trees with method of phylogenetic inference and bootstrap values used to assess reassortant status were high, regardless of method of inference.

The disparate segment-specific phylogenies of viruses in the 1991–2000 reference trees suggested that reassortment occurred rapidly compared with the time since the ancestor of contemporary avian influenza A viruses. Our estimate of the rate of reassortment among segments of the replication complex of viruses from Eurasian domestic aquatic and terrestrial avian hosts led to the conservative estimate of 3 years for the median duration between creation of a new genotype and its next segment reassortment (this estimate is additionally conservative because the rate of reassortment between significantly diverged and successfully replicating viruses, which we estimated here, does not account for reassortment events between more closely related viruses or reassortment events that have a deleterious effect on the viral fitness). The frequency of reassortment demonstrates that dual infection of avian species is not an uncommon event and it frequently results in the generation of reassortants with unpredictable phenotype. Supporting our conclusion that influenza viruses infecting domestic avian species undergo frequent reassortment are studies identifying a number of genotypes of avian H5N1 virus in South-East Asia between 1999 and 2003, each circulating for no more than 2 years, with the eventual dominance of the Z genotype (Li *et al.*, 2004).

Our analyses identified novel lineages of PB2 and PA introduced by reassortment into the influenza viruses infecting domestic birds; we speculate that wild aquatic birds, the natural

hosts of influenza A viruses, are the source of this newly introduced material. Evidence for genomes that include novel genetic material has been reported (Choi *et al.*, 2004; Reid *et al.*, 2004) and we see in our study that this is not an uncommon occurrence. This is of concern, considering the vast diversity of influenza viruses within wild aquatic birds and the potential for domestic avian species to transmit viruses to humans and swine.

It is intriguing to speculate on whether selection plays a role in genotype turnover in avian hosts. In humans, NP, which contains epitopes for cytotoxic T-cell activity in humans, has been shown to evolve to escape this immune pressure (Boon *et al.*, 2002); a similar mechanism might be at work in avian species. The rapid rate of reassortment measured here has serious implications for the potential to generate a pandemic strain: reassortment creates evolutionary change at a faster rate than point mutations in a non-reassortant virus, possibly increasing the rate of reaching a viral genomic configuration capable of reassorting with human-adapted viruses. As reassortment is an opportunistic process of change (as distinct from the unavoidable process of nucleotide change due to RNA replication), its rate depends on the context for viral population growth, involving such factors as host species and agricultural practices. If livestock-management practices are changed to reduce the potential for reassortment, our procedure for estimating the rate of reassortment, applied in real time, may provide a quantitative measure of the success of these changes.

ACKNOWLEDGEMENTS

We thank David D. Pollock for fruitful discussions. We are grateful to the two programmers of the ISD, Rachel Richard and David McDonald, for excellent technical assistance. Financial support was provided to C. A. M. and W. J. B. by NIH grant no. AI 95357 and to R. J. W. by ALSAC (American Lebanese Syrian Associated Charities).

REFERENCES

- Alexander, D. J. (2000).** A review of avian influenza in different bird species. *Vet Microbiol* **74**, 3–13.
- Area, E., Martín-Benito, J., Gastaminza, P., Torreira, E., Valpuesta, J. M., Carrascosa, J. L. & Ortín, J. (2004).** 3D structure of the influenza virus polymerase complex: localization of subunit domains. *Proc Natl Acad Sci U S A* **101**, 308–313.
- Boon, A. C. M., de Mutsert, G., Graus, Y. M. F., Fouchier, R. A. M., Sintnicolaas, K., Osterhaus, A. D. M. E. & Rimmelzwaan, G. F. (2002).** Sequence variation in a newly identified HLA-B35-restricted epitope in the influenza A virus nucleoprotein associated with escape from cytotoxic T lymphocytes. *J Virol* **76**, 2567–2572.
- Bruno, W. J., Socci, N. D. & Halpern, A. L. (2000).** Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol Biol Evol* **17**, 189–197.
- Chare, E. R., Gould, E. A. & Holmes, E. C. (2003).** Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. *J Gen Virol* **84**, 2691–2703.
- Choi, Y. K., Ozaki, H., Webby, R. J., Webster, R. G., Peiris, J. S., Poon, L., Butt, C., Leung, Y. H. C. & Guan, Y. (2004).** Continuing evolution of H9N2 influenza viruses in Southeastern China. *J Virol* **78**, 8609–8614.
- Clements, M. L., Subbarao, E. K., Fries, L. F., Karron, R. A., London, W. T. & Murphy, B. R. (1992).** Use of single-gene reassortant viruses to study the role of avian influenza A virus genes in attenuation of wild-type human influenza A virus for squirrel monkeys and adult human volunteers. *J Clin Microbiol* **30**, 655–662.
- Dorman, K. S., Kaplan, A. H. & Sinsheimer, J. S. (2002).** Bootstrap confidence levels for HIV-1 recombination. *J Mol Evol* **54**, 200–209.
- Felsenstein, J. (1993).** PHYLIP (phylogeny inference package), version 3.5c. Department of Genome Sciences, University of Washington, Seattle, USA.
- Guan, Y., Shortridge, K. F., Krauss, S. & Webster, R. G. (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**, 9363–9367.
- Hasegawa, M., Kishino, H. & Yano, T. (1985).** Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**, 160–174.
- Hatchette, T. F., Walker, D., Johnson, C., Baker, A., Pryor, S. P. & Webster, R. G. (2004).** Influenza A viruses in feral Canadian ducks: extensive reassortment in nature. *J Gen Virol* **85**, 2327–2337.
- Hatta, M., Gao, P., Halfmann, P. & Kawaoka, Y. (2001).** Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science* **293**, 1840–1842.
- Hiromoto, Y., Yamazaki, Y., Fukushima, T. & 7 other authors (2000).** Evolutionary characterization of the six internal genes of H5N1 human influenza A virus. *J Gen Virol* **81**, 1293–1303.
- Hoffmann, E., Stech, J., Leneva, I., Krauss, S., Scholtissek, C., Chin, P. S., Peiris, M., Shortridge, K. F. & Webster, R. G. (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**, 6309–6315.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S. & Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**, 1789–1796.

Li, K. S., Guan, Y., Wang, J. & 19 other authors (2004). Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature* **430**, 209–213.

Macken, C. A., Lu, H., Goodman, J. & Boykin, L. (2001). The value of a database in surveillance and vaccine selection. In *Options for the Control of Influenza IV*, p. 103–106. Edited by A. Osterhaus, N. Cox & A. Hampson. Amsterdam: Elsevier Science.

Murphy, B. R., Buckler-White, A. J., London, W. T., Harper, J., Tierney, E. L., Miller, N. T., Reck, L. J., Chanock, R. M. & Hinshaw, V. S. (1984). Avian-human reassortant influenza A viruses derived by mating avian and human influenza A viruses. *J Infect Dis* **150**, 841–850.

Obenauer, J. C., Denson, J., Mehta, P. K. & 14 other authors (2006). Large-scale sequence analysis of avian influenza isolates. *Science* **311**, 1576–1580.

Rambaut, A. (2000). Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**, 395–399.

Reid, A. H., Taubenberger, J. K. & Fanning, T. G. (2004). Evidence of an absence: the genetic origins of the 1918 pandemic influenza virus. *Nat Rev Microbiol* **2**, 909–914.

Scholtissek, C., Stech, J., Krauss, S. & Webster, R. G. (2002). Cooperation between the hemagglutinin of avian viruses and the matrix protein of human influenza A viruses. *J Virol* **76**, 1781–1786.

Snyder, M. H., Buckler-White, A. J., London, W. T., Tierney, E. L. & Murphy, B. R. (1987). The avian influenza virus nucleoprotein gene and a specific constellation of avian and human virus polymerase genes each specify attenuation of avian-human influenza A/Pintail/79 reassortant viruses for monkeys. *J Virol* **61**, 2857–2863.

Swofford, D. (1993). PAUP*: phylogenetic analysis using parsimony. Champaign: University of Illinois.

Wagner, R., Matrosovich, M. & Klenk, H.-D. (2006). Functional balance between haemagglutinin and neuraminidase in influenza virus infections. *Rev Med Virol* **12**, 159–166.

Wallensten, A., Munster, V. J., Elmberg, J., Osterhaus, A. D., Fouchier, R. A. & Olsen, B. (2005). Multiple gene segment reassortment between Eurasian and American lineages of influenza A virus (H6N2) in Guillemot (*Uria aalge*). *Arch Virol* **150**, 1685–1692.

Wilson, A. C., Carlson, S. S. & White, T. J. (1977). Biochemical evolution. *Annu Rev Biochem* **46**, 573–639.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555–556.

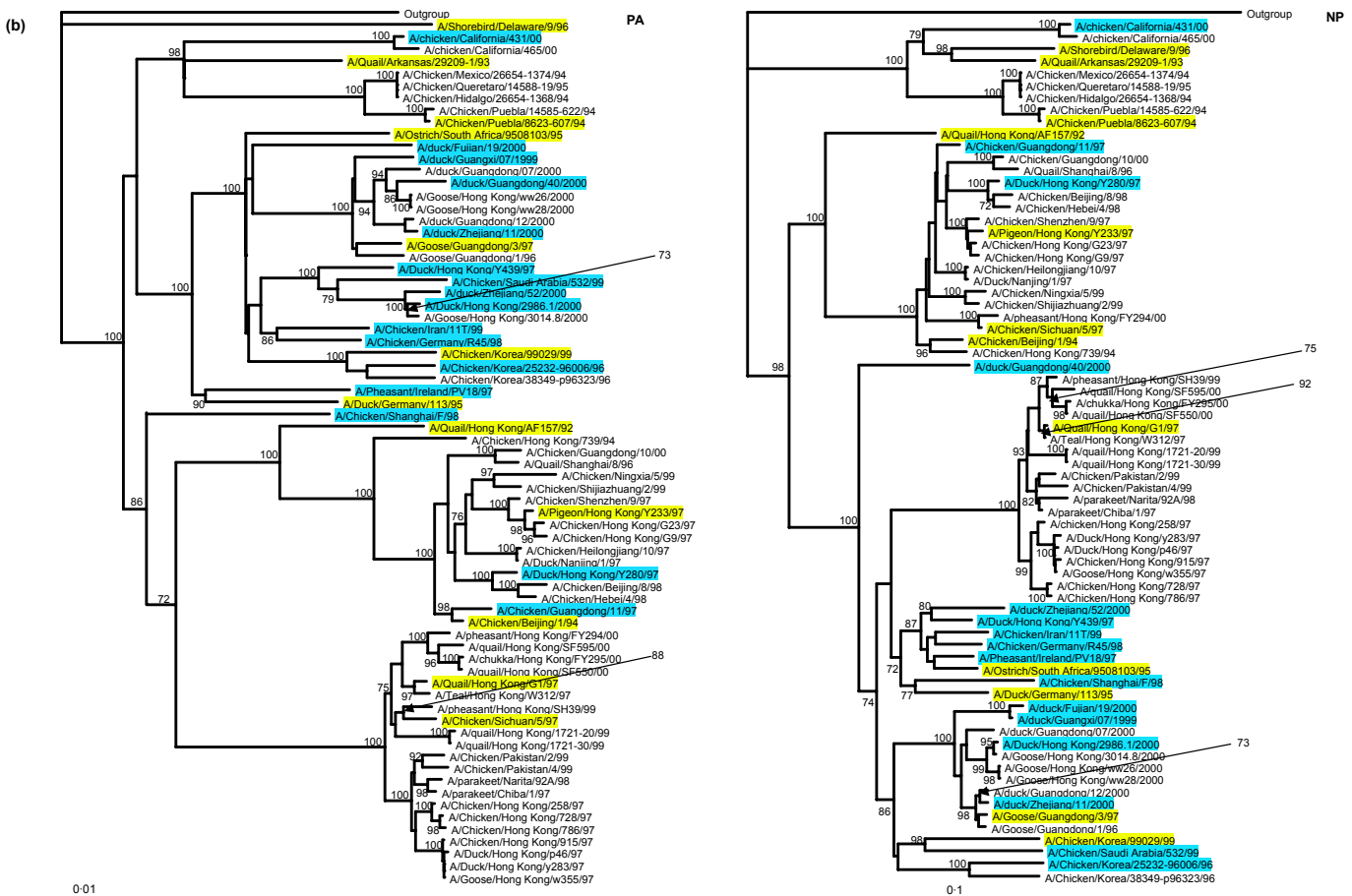


Fig. 1. Segment-specific phylogenies for the replication complex of 1991–2000 (T1) avian influenza A viruses with approximately full-length PB2, PB1, PA and NP sequences; numbers give bootstrap support as a percentage. Highlighted viruses (yellow or blue) comprise the reference set, the collection of single representatives of each distinct genotype present in the 1991–2000 dataset. Yellow viruses are contained in the reference trees (Fig. 2); blue viruses do not appear in the reference trees because they are attached with low bootstrap support at a deep node in at least one reference tree.

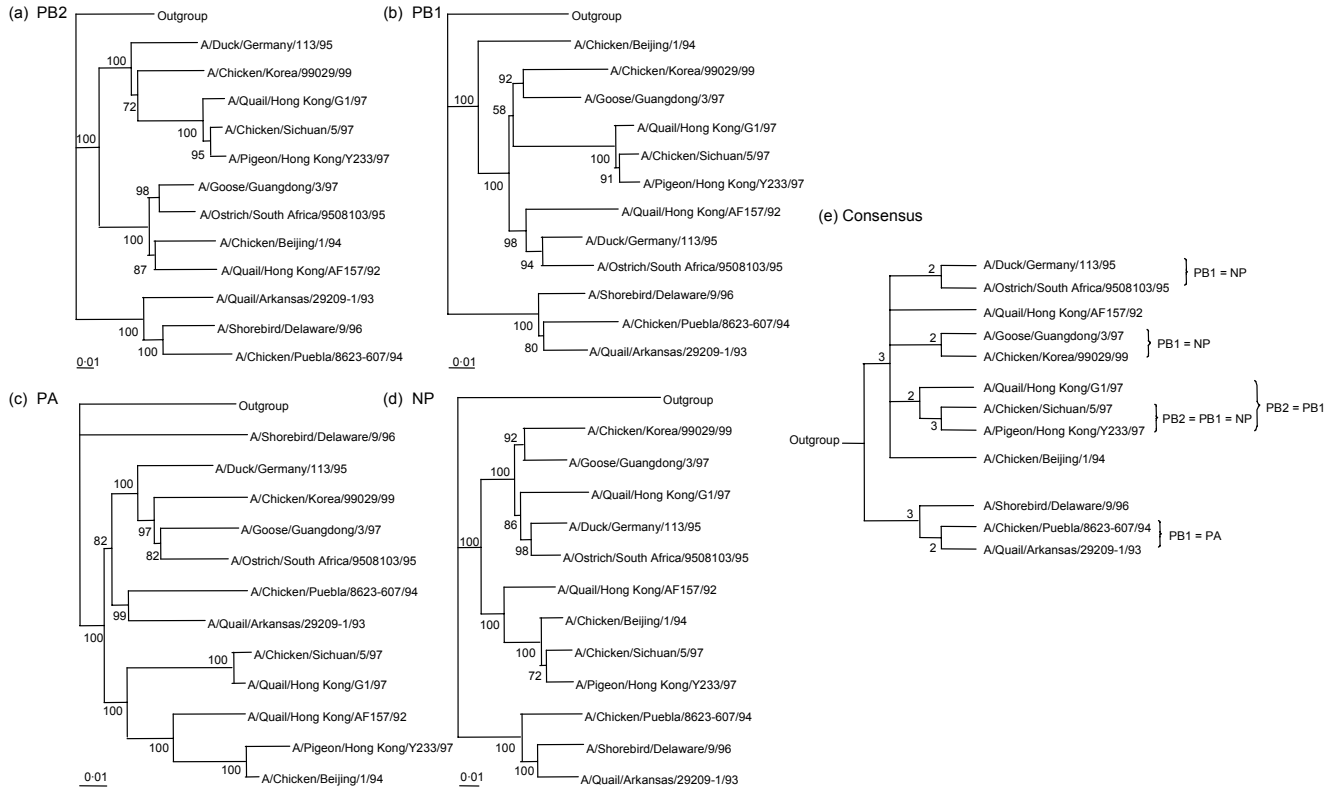


Fig. 2. Segment-specific reference trees of viruses used to assess the reassortant status of the replication complex of 2001–2004 (T2) viruses and the consensus of these trees. (a–d) Reference trees containing those viruses highlighted in yellow in Fig. 1. Numbers give bootstrap support as a percentage. (e) Consensus of (a–d). A number at a node gives the count of segments for which the viruses in the subtree at this node group together (branching patterns appearing less than twice have been collapsed). Annotation such as ‘PB2=PB1’ denotes the segments for which the viruses in the subtree group together.

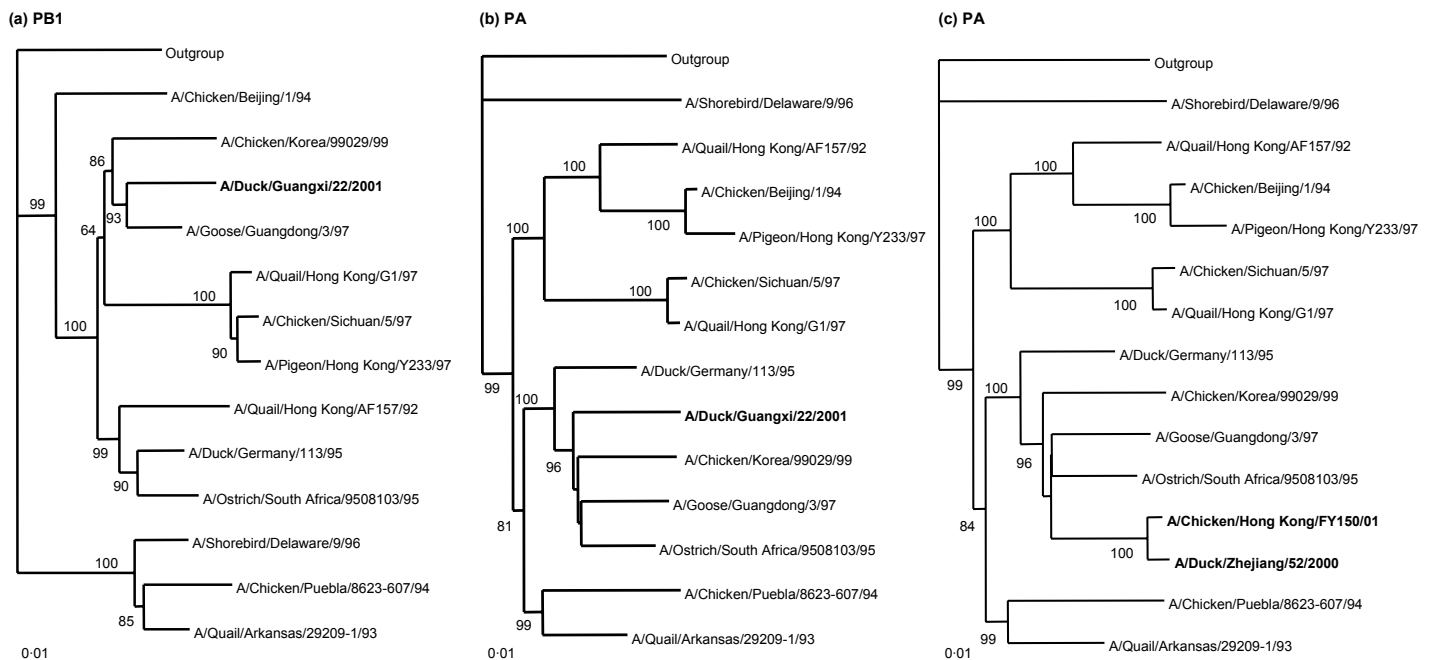


Fig. 3. Examples of placement of T2 viruses on the T1 reference trees. In the majority of cases, when a segment of a T2 virus was added to the respective segment-specific T1 reference tree (Figs 2a–d), it associated with a single reference-tree virus. For example, in (a), the PB1 of *A/duck/Guangxi/22/2001* associated with *A/Goose/Guangdong/3/97*. Sometimes, a T2 viral segment attached basal to a subtree of the respective T1 reference tree, as, for example, did the PA of *A/duck/Guangxi/22/2001* in (b). Basal attachments usually indicated a T2 segment that matched a segment of a virus in the T1 reference set, but not in the T1 reference tree. For example, in (c), *A/chicken/Hong Kong/FY150/2001* had the same basal attachment as *A/duck/Zhejiang/52/2000*, a virus in the reference set, but not in the reference tree. Numbers give bootstrap support as a percentage.

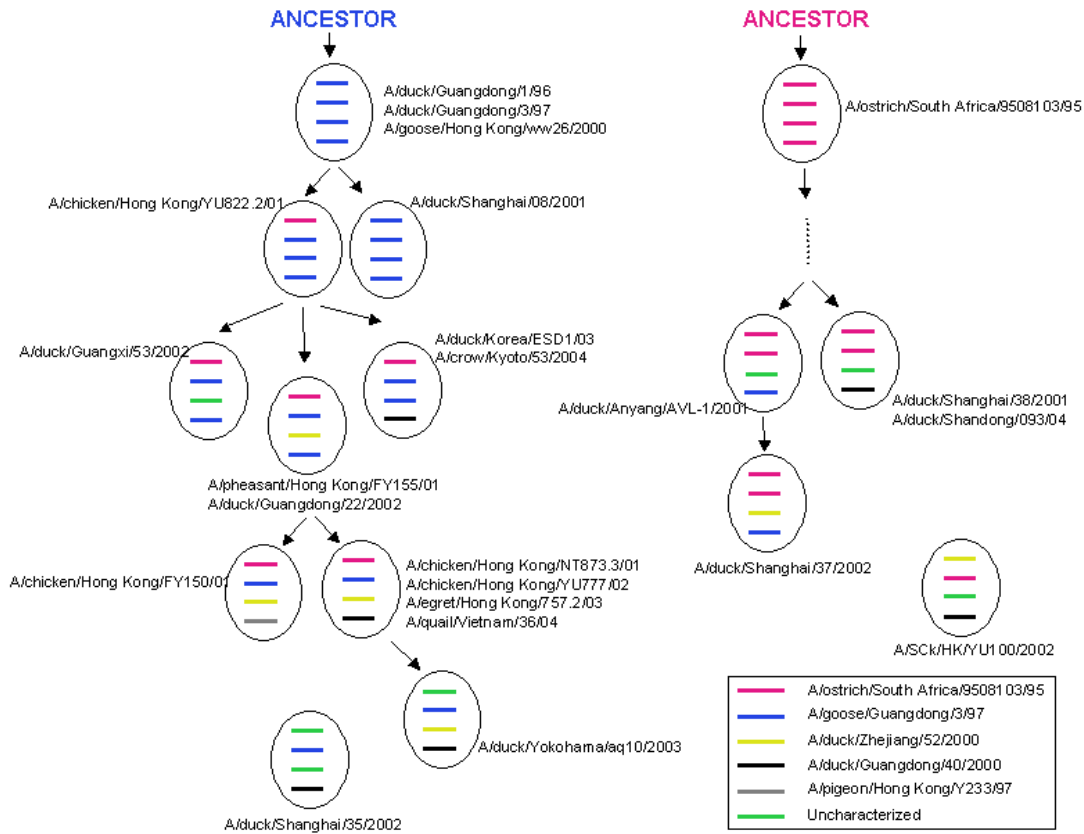


Fig. 4. One of three maximum-parsimony relationships among genotypes of the replication complex of influenza A viruses from Eurasian terrestrial avian hosts sampled between 2001 and 2004. Relationships among the genotypes result from reassorting single segments in the sequence shown. Reassortment of one segment at a time has strong support from the two-time test of the reassortant status of the replication complex of these viruses. Coloured bars represent, in order: PB2 (top), PB1, PA and NP; colour represents the donor virus in the reference set. One virus from each year that a genotype was observed is illustrated.

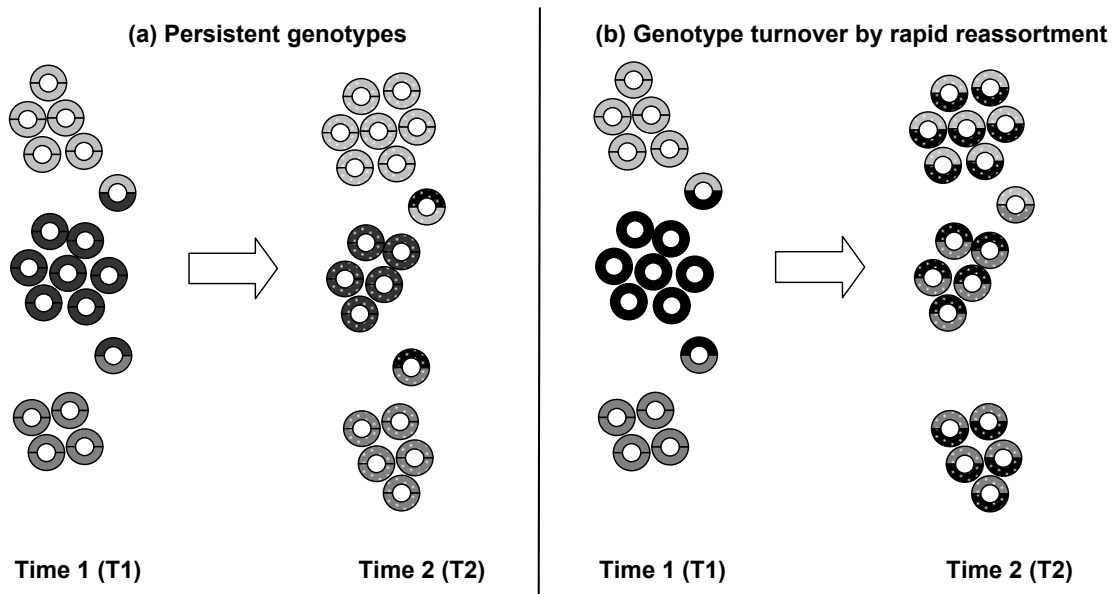


Fig. 5. Two possible scenarios for the effect of reassortment on the evolution of two segments (upper and lower arcs) of *Influenza A virus*. Shades of grey, lineages; stipples, evolution by point mutation. Without prior history, the reassortant status of T1 viruses cannot be determined and both segments are given the same shade. (a) Non-reassortants have a fitness advantage; segment-specific phylogenies look the same with the exception of a rare reassortant. This case is often referred to as ‘preserving a clonal frame’. (b) Reassortants are generated frequently; reassortant viruses replace earlier viruses; segment-specific phylogenies differ widely (the clonal frame is not preserved). The two-time test applied to avian influenza A viruses from 1978–2004 from both the North American and Eurasian lineages points to the four-segment analogue of scenario (b) being the more realistic representation of the effect of reassortment on the replication complex of these viruses.

Table 1. Reassortant status of Asian viruses from 2001–2004 tested against the 1991–2000 reference set

Abbreviations: GsGu, A/Goose/Guangdong/3/97; Ostr, A/Ostrich/South Africa/9508103/95; DkGer, A/Duck/Germany/113/95; Pg, A/Pigeon/Hong Kong/Y233/97; QuG1, A/Quail/Hong Kong/G1/97; CkBeij, A/Chicken/Beijing/1/94; CkSich, A/Chicken/Sichuan/5/97; QuAF, A/Quail/Hong Kong/AF157/92; Shore, A/shorebird/Delaware/9/96; QuArk, A/quail/Arkansas/29209-1/93; CkPueb, A/chicken/Puebla/8623-607/94; CkKor, A/chicken/Korea/99029/99.

Strain of T2 virus	Placement of candidate reassortants on segment-specific T1 reference trees*			
	PB2	PB1	PA	NP
Asian:				
A/Duck/Shanghai/08/2001	GsGu	GsGu	GsGu	GsGu
A/Chicken/Hong Kong/YU822.2/01	Ostr	GsGu	GsGu	GsGu
A/Duck/Guangxi/35/2001	Ostr	GsGu	GsGu	GsGu
A/Duck/Guangxi/22/2001	Ostr	GsGu†	(CkKor,GsGu,Ostr)†	GsGu
A/Pheasant/Hong Kong/FY155/01	Ostr	GsGu	(CkKor,GsGu,Ostr)	GsGu
A/Chicken/Hong Kong/FY150/01	Ostr	GsGu	(CkKor,GsGu,Ostr)	Pg
A/Chicken/Hong Kong/NT873.3/01	Ostr	GsGu	(CkKor,GsGu,Ostr)	(DkGer,Ostr,QuG1,CkKor,GsGu)
A/Duck/Shanghai/38/2001	Ostr	Ostr	(CkSich,QuG1,Pg,CkBeij,QuAF)	(DkGer,Ostr,QuG1,CkKor,GsGu)
A/Duck/Anyang/AVL-1/2001	Ostr	Ostr	(CkSich,QuG1,Pg,CkBeij,QuAF)	GsGu
A/Duck/Guangxi/53/2002	Ostr	GsGu	(CkSich,QuG1,Pg,CkBeij,QuAF)	GsGu
A/Chicken/Hong Kong/YU777/02	Ostr	GsGu	(CkKor,GsGu,Ostr)	(DkGer,Ostr,QuG1,CkKor,GsGu)
A/Duck/Shanghai/37/2002	Ostr	Ostr	(CkKor,GsGu,Ostr)	GsGu
A/Duck/Guangdong/22/2002	Ostr	GsGu	(CkKor,GsGu,Ostr)	GsGu
A/Duck/Fujian/13/2002	Ostr	Ostr	(CkKor,GsGu,Ostr)	GsGu
A/Duck/Shanghai/35/2002	(GsGu,Ostr,CkBeij,QuAF)	GsGu	(CkSich,QuG1,Pg,CkBeij,QuAF)	(DkGer,Ostr,QuG1,CkKor,GsGu)
A/chicken/Hong Kong/31.4/02	Ostr	GsGu	(CkKor,GsGu,Ostr)	GsGu
A/SCK/HK/YU100/2002	(Ostr,GsGu)	Ostr	(CkKor,GsGu,DkGer,Ostr)	(DkGer,Ostr,QuG1,CkKor,GsGu)

Strain of T2 virus	Placement of candidate reassortants on segment-specific T1 reference trees*			
	PB2	PB1	PA	NP
A/duck/Yokohama/aq10/2003	DkGer	GsGu	(CkKor, GsGu, Ostr)	(DkGer, Ostr, QuG1, CkKor, GsGu)
A/Chicken/Jilin/9/2004	Ostr	GsGu	(CkKor, GsGu, Ostr)	(DkGer, Ostr, QuG1, CkKor, GsGu)
A/Dk/Indonesia/MS/2004	Ostr	GsGu	(CkKor, GsGu, Ostr)	(DkGer, Ostr, QuG1, CkKor, GsGu)
A/chicken/Guangdong/191/04	Ostr	GsGu	(CkKor, GsGu, Ostr)	(DkGer, Ostr, QuG1, CkKor, GsGu)
A/quail/Vietnam/36/04	Ostr	GsGu	(CkKor, GsGu, Ostr)	(DkGer, Ostr, QuG1, CkKor, GsGu)
A/duck/Guangdong/173/04	Ostr	GsGu	(CkKor, GsGu, Ostr)	(DkGer, Ostr, QuG1, CkKor, GsGu)
A/crow/Kyoto/53/2004	Ostr	GsGu	GsGu	(DkGer, Ostr, QuG1, CkKor, GsGu)
A/Chicken/Guangdong/174/04	Ostr	GsGu	GsGu	(DkGer, Ostr, QuG1, CkKor, GsGu)
A/duck/Shandong/093/2004	Ostr	Ostr	(CkSich, QuG1, Pg, CkBeij, QuAF)	(DkGer, Ostr, QuG1, CkKor, GsGu)
European/American:				
A/Chicken/Netherlands/1/03	DkGer	(DkGer, Ostr)	GsGu	(DkGer, Ostr, QuG1, CkKor, GsGu)
A/Duck/NC/91347/01	Shore	(Shore, CkPueb, QuArk)	QuArk	Basal to all
A/Chicken/California/139/01	QuArk	(Shore, CkPueb, QuArk)	(CkPueb, QuArk)	(Shore, QuArk, CkPueb)
A/Chicken/Chile/176822/02	(Shore, CkPueb, QuArk)	(Shore, CkPueb, QuArk)	Basal to all	Basal to all
A/Chicken/British Columbia/04	QuArk	(Shore, CkPueb, QuArk)	(CkPueb, QuArk)	(Shore, QuArk)

*When a single virus is named, attachment occurred with high bootstrap support for being the nearest neighbour to that virus. When multiple viruses are named, attachment occurred with medium to high bootstrap support basal to those viruses. Bold type denotes a segment of an Asian virus that is not represented in the 1991–2000 reference set.

†Attachment of segment to the T1 reference tree is illustrated in Fig. 3.