*D376–D380* *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 25 October 2006*
*doi:10.1093/nar/gkl779*

# Influenza Virus Database (IVDB): an integrated information resource and analysis platform for influenza virus research

**Suhua Chang[1,2], Jiajie Zhang[1], Xiaoyun Liao[1], Xinxing Zhu[1,3], Dahai Wang[1], Jiang Zhu[1], Tao Feng[1], Baoli Zhu[4,5], George F. Gao[4,5], Jian Wang[1], Huanming Yang[1,4], Jun Yu[1,*] and Jing Wang[1,4,*]**

[1]Beijing Institute of Genomics, Chinese Academy of Sciences, B-6 Airport Industrial Zone, Beijing 101300, China, [2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China, [3]James D. Watson Institute of Genome Sciences of Zhejiang University, Hangzhou 310007, China, [4]Joint Center for Microbial Genomics, Chinese Academy of Sciences, Beijing 101300, China and [5]Center for Molecular Virology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100080, China

Received August 12, 2006; Revised September 18, 2006; Accepted October 1, 2006

## ABSTRACT

**Frequent outbreaks of highly pathogenic avian influenza and the increasing data available for comparative analysis require a central database specialized in influenza viruses (IVs). We have established the Influenza Virus Database (IVDB) to integrate information and create an analysis platform for genetic, genomic, and phylogenetic studies of the virus. IVDB hosts complete genome sequences of influenza A virus generated by Beijing Institute of Genomics (BIG) and curates all other published IV sequences after expert annotation. Our Q-Filter system classifies and ranks all nucleotide sequences into seven categories according to sequence content and integrity. IVDB provides a series of tools and viewers for comparative analysis of the viral genomes, genes, genetic polymorphisms and phylogenetic relationships. A search system has been developed for users to retrieve a combination of different data types by setting search options. To facilitate analysis of global viral transmission and evolution, the IV Sequence Distribution Tool (IVDT) has been developed to display the worldwide geographic distribution of chosen viral genotypes and to couple genomic data with epidemiological data. The BLAST, multiple sequence alignment and phylogenetic analysis tools were integrated for online data analysis. Furthermore, IVDB offers instant access to pre-computed alignments and polymorphisms of IV genes and proteins, and presents the results as SNP distribution plots and minor allele distributions. IVDB is publicly available at http://influenza.genomics.org.cn**

## INTRODUCTION

Influenza virus (IV) is the causative agent of several serious influenza pandemics (1). Recently, a highly pathogenic avian influenza virus (AIV; H5N1) has resulted in the death of more than 100 people and the slaughter of millions of poultry in Asia, Europe and Africa (World Health Organization, http://www.who.int). Scrupulous surveillance and multidisciplinary interrogation of the viral migration patterns and evolution are crucial for preventing further casualties of humans and domestic poultry. Viral genome sequences provide essential information for understanding pathogenesis, diagnosis, and therapy of the virus. Since the IV genome mutates very fast from host to host and from year to year, tracing IV lineages and discovering the pattern of sequence variation provide a solid foundation for evolutionary and functional studies (2). Genome-wide sequence analyses of IV have demonstrated various genotypes and proteotypes (3) and revealed multiple lineages and genetic re-assortment among viruses (4,5).

The Beijing Institute of Genomics (BIG) has been sequencing IVs collected by scientists from different institutions from different parts of China. Since a highly pathogenic avian influenza A subtype H5N1 virus [A/Goose/Guangdong/1/96/(H5N1)] was initially isolated from China and subsequently identified as a precursor of the Hong Kong 1997 AIV (A/Hong Kong/156/97) (6), China, especially the southern

*To whom correspondence should be addressed. Tel: +86 10 80485492; Fax: +86 10 80498676; Email: wangjing@genomics.org.cn
*Correspondence may also be addressed to Jun Yu. Tel: +86 10 80481455; Fax: +86 10 80498676; Email: junyu@genomics.org.cn

The authors wish it to be known that, in their opinion, the first three authors should be regarded as First Authors.

provinces, became one of the major foci for viral surveillance. Phylogenetic analysis of all gene segments of H5N1 viruses isolated in China and other countries allowed us to trace the ecological and genetic origins of AIV (7) and the mechanism of its transmission to Southeast Asia, Europe and Africa (8). We have sequenced isolates of AIV subtype H5N1 from 1997 to 2005 in different hosts found in China, such as wild birds, poultry, water fowl and mammals. Some of these strains have been published, such as those from tree sparrows in Henan Province in 2004 (9) and migratory birds in Qinghai Province in 2005 (10).

Significant international efforts in sequencing viral isolates have been made worldwide, so the number of IV sequences has been rising rapidly in public resources, such as the Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) at the National Center for Biotechnology Information (NCBI) and the Influenza Sequence Database (ISD) at Los Alamos National Laboratory (11). The NCBI Influenza Virus Resource presents data obtained from the NIAID (National Institute of Allergy and Infectious Diseases) Influenza Genome Sequencing Project as well as from GenBank. Since most of the IV sequences are direct submissions, there is a growing need to curate and corroborate the data in order to analyze them effectively. ISD provides curated data, but it requires a subscription fee to obtain full access to the data and tools. Therefore, to provide a central IV-specialized database, which curates IV sequences, integrates information, and provides free online tools for data analysis, is of great importance. The Influenza Virus Database (IVDB) contains both BIG's data and published IV sequences after expert curation to ensure a high standard of accuracy and completeness. We have further developed tools and viewers to analyze and browse our data that include information concerning genomes, genes, polymorphisms, and phylogenetic relationships. IVDB aims to be a powerful information resource and an analysis workbench for scientists working on IV genetics, evolution, diagnostics, vaccine development and drug design.

## DATA CONTENT AND DATA CURATION

IVDB contains both BIG's data (BIG has an on-going effort to generate more AIV/IV sequences) and data from public resources. Data from NCBI's Influenza Virus Resource form the backbone of this database. Since ISD accepts direct submissions and contains some IV sequences that are absent from NCBI's Influenza Virus Resource, IVDB incorporates an additional 1654 segment sequences from ISD. The data types hosted in IVDB include viral type/subtype, source information, nucleotide/protein sequence [classified as complete genome, CDS, untranslated region (UTR), three-dimensional (3-D) structural data], sequence alignment, polymorphism, and categorized literature and web resources covering IV genomics, pathogenicity and epidemiology. The current protein 3-D data are mainly from Protein Data Bank (12). More 3-D predictions will be acquired from our collaborators in protein 3-D modeling.

All data in IVDB have been curated manually since data quality is of crucial importance for analysis. To ensure adequate data quality, we first examine annotations and source information from each sequence entry in public databases. If we see inconsistencies or errors in the records, such as serotypes, we double check and proof-read the sequences. Sequences that do not have a genotype and subtype assigned are manually typed using phylogenetic analysis and BLAST tools. Second, we redefine host species (e.g. goose, coot instead of avian in general) and sampling locations (not only continents and countries/regions, but also provinces/states), and store the information in searchable fields. The detailed host information is crucial since phylogenetic studies of influenza A virus isolates have revealed that the viral genes form species-specific lineages (13). This is especially useful for AIV research since users have to distinguish wild birds from poultry, particularly the aquatic birds that are thought to be the primary carriers of influenza A viruses (13). Third, we match protein sequences with nucleotide sequences, and further annotate some predicted CDS and UTRs. Users may browse the nucleotide and protein information simultaneously in one single record and trace back to the original records, if available, through hyperlinks. Furthermore, we developed the IV Sequence Quality Filter System (Q-Filter) that classifies the nucleotide sequences into 7 categories of C1–C4 and P1–P3, respectively, according to their sequence content (CDS, 5′-UTR, 3′-UTR) and integrity (C: complete or P: partial). The C1–C4 sequences have complete CDS and differ in UTRs (such as without UTR, with 5′/3′-UTR or both). The P1–P3 sequences have partial CDS associated with no UTR or with 5′/3′-UTR. All nucleotide sequences in IVDB have been classified and users may choose a specific category of data for analysis. It is often useful to filter out short sequences since the length of nucleotide sequences ranges from dozens to thousands of base pairs. This allows researchers to have a clear understanding of the sequence data and to choose high-quality data for their research.

The website also provides pre-computed alignments, phylogenetic trees, and variations in IV genes and proteins, which are grouped by host, subtype and segment. The aligned sequence groups have been manually corrected to remove redundant sequences. Results are presented through a graphical view of SNP distribution or minor allele distribution, as well as tabular statistics on each nucleotide position compared with the consensus sequence. Users are not only able to search for sequence polymorphisms by host, subtype, and segment but also have instant access to pre-made alignments, phylogenetic trees, and geographical distributions on a world map.

The current version of IVDB contains 35 549 IV nucleotide sequences, most of which are from influenza A virus with 5670 H5N1 sequences. Nearly 2400 and 700 are from influenza B and C viruses, respectively. There are 2687 IV genome sets that contain all eight segments of type A and B IV, or all seven segments of type C IV. More than 40 000 protein sequences, 118 protein 3-D structure records, and ~200 groups of multiple sequence alignments with polymorphic sites highlighted are also included. The statistics of nucleotide sequences available on September 16, 2006 are shown in Table 1. More detailed statistical figures and tables are available on the IVDB website. All data are freely available for downloading from our FTP site (ftp://ftp.genomics.org.cn/pub/influenza/).

**Table 1.** Statistics of nucleotide sequences in IVDB (September 16, 2006)

| Nucleotide sequence | Type A[a] | Type B[a] | Type C[a] | Subtype H5N1 | Total |
|---|---|---|---|---|---|
| All segments | 32 537 | 2317 | 695 | 5670 | 35 549 |
| PB2 | 3170 | 92 | 90 | 622 | 3352 |
| PB1 | 3150 | 94 | 88 | 623 | 3332 |
| PA(P3[b]) | 3102 | 95 | 88 | 625 | 3285 |
| HA(HE[b]) | 7473 | 1265 | 114 | 942 | 8852 |
| NP | 3715 | 101 | 87 | 641 | 3903 |
| NA | 4275 | 402 | / | 861 | 4677 |
| M | 3894 | 98 | 107 | 693 | 4099 |
| NS | 3758 | 170 | 121 | 663 | 4049 |
| GenomeSet | 2551 | 72 | 64 | 458 | 2687 |

[a]The total number of nucleotide sequences from influenza A, B and C viruses.
[b]The segment name of influenza C virus.

## DATABASE ACCESS AND TOOLKIT

IVDB provides a powerful Search Engine for users to retrieve different data types hosted in IVDB. Except for keyword search, users may combine various search options such as type (influenza type A/B/C virus), serotype, host species, sampling place, sampling year, sequence length range, and sequence category classified by Q-Filter. If users select the GenomeSet option, IVDB will display complete sets of sequence segments from selected IV strains. A key enhancement is to allow users to browse results from the first search and to acquire more targeted data sets through a second search. For example, users can acquire full-length HA sequences of H5N1 serotype isolated from chickens in Guangdong Province of China between 2003 and 2005. All search results can be saved to a personalized WorkSet for

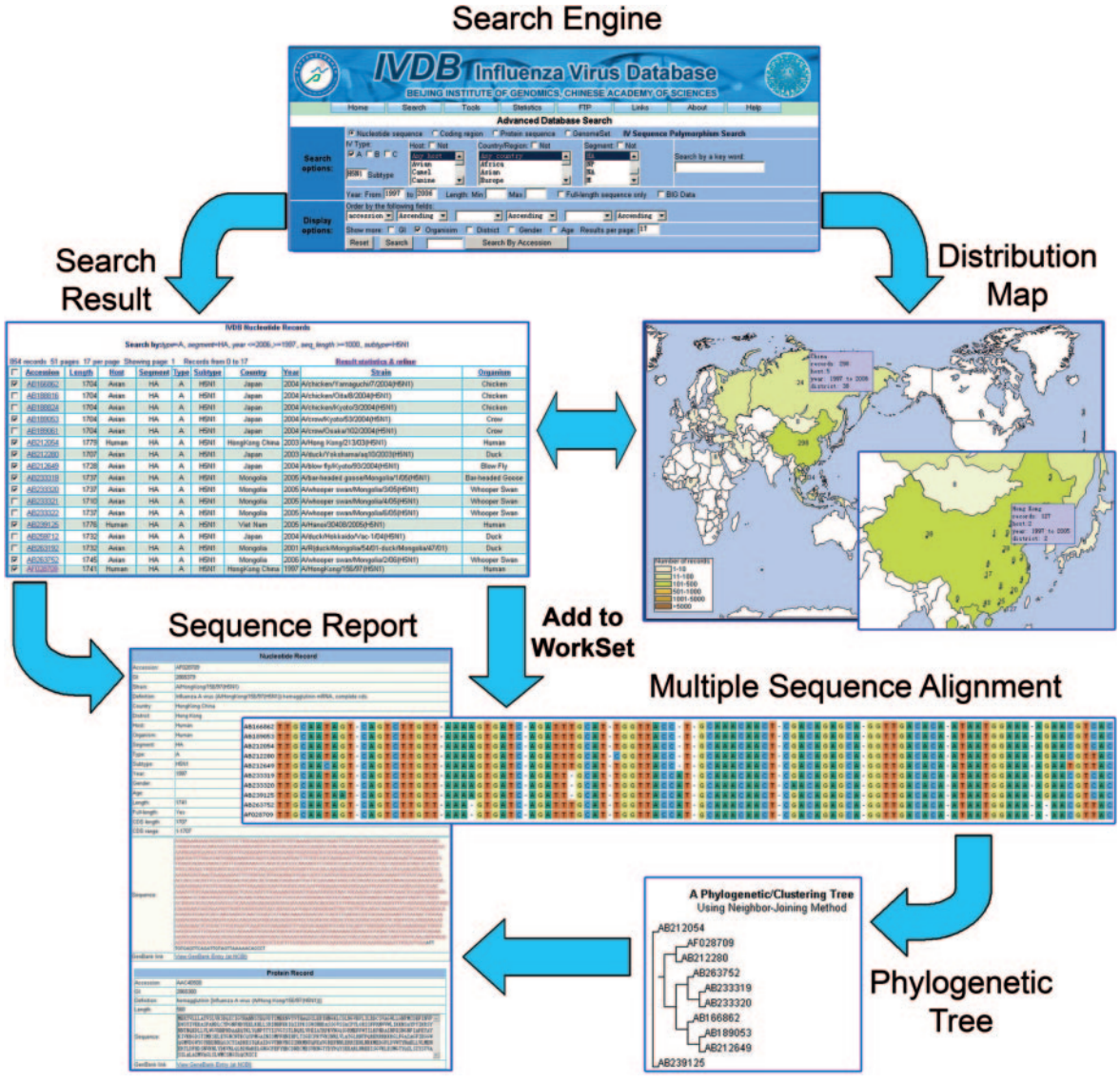


**Figure 1.** Screenshot showing the interrelation of IVDB data and tools. Users access the data through Search Engine. Search results can be selectively saved in a personalized WorkSet and subjected to successive data analyses, such as plotting geographical distribution, aligning multiple sequences, and building phylogenetic trees.

users' convenience in data management. Users can also maintain the WorkSet themselves, delete, append or download data as well as apply successive analyses to the data using the tools provided.

As an IV data analysis platform, IVDB provides a toolkit and a series of viewers for analyzing IV genomes, genes, polymorphisms, and phylogenetic relationship individually or in a comparative context. IVDB integrates BLAST tools for comparing user-supplied data (e.g. WorkSet) with IVDB's built-in databases, ClustalW for multiple sequence alignment (14), MUSCLE (15) and PHYLIP (16) for building phylogenetic trees. To facilitate analysis of global viral transmission and evolution, we developed the IV Sequence Distribution Tool (IVDT) for plotting worldwide geographic distributions of IV sequences. IVDT draws maps indicating geographical origins and the frequency of our database records. It displays a zoomed-in geographic distribution of IV sequences from an upper level world map to a bottom-level of nation map. When the cursor is placed on a specific area of interest, it will show sequence statistics in the area with information on hosts and years of sampling. When the user clicks on the bottom-level map, it will display a list of hyperlinked sequence records that link to more details. Users may alternate easily from their search results to sequence distribution maps, and vice versa. IVDT facilitates the coupling of sequence data with geographical information and epidemiological data. The IVDT's sequence distribution map is a useful visual aid to display in which countries and at what density the samples have been collected in different regions of the world. However, its results need to be interpreted with care since sampling biases are likely to be present.

All data and tools housed in IVDB are cross-linked, and the whole IVDB website forms an interrelated network as illustrated in Figure 1, which helps users to utilize, analyze, and understand the data more effectively and efficiently.

## SYSTEM DESIGN AND IMPLEMENTATION

IVDB was developed using our established pipeline for biological databases (17–19). It consists of three hardware components, a World Wide Web server, a database server, and a server for sequence analysis. The system is based on a MySQL relational database, and the front end consists of a set of JSP scripts running on a TomCat web server. The Q-Filter and search engine were developed using Java. IVDT is similar to GIS (Geographic Information System) with country coordinates stored in the database. The system searches the coordinate data and draws the distribution map after the search. The BLAST and multiple sequence alignment tools run on clusters of super computers, and computational tasks are submitted by PBS (Portable Batch System).

## FUTURE DEVELOPMENT

We are aiming to build a central IV-specialized database functioning not only as an integrated information resource, but also as an analysis platform for genetic, genomic, and phylogenetic research of IVs. Continuing efforts will be made to update IV sequences and incorporate new data and resources as soon as they become available. IVDB will further classify IV sequences into host-specific clusters since IVs are isolated from a broad range of species. For analyzing transmission patterns in wild birds, IVDB will offer additional information on the geohydrologic environment, seasonal migrations, and migratory flyways of recorded wild bird populations throughout the world (20), providing useful references for sequence variations from IVDB. In-depth IV mutation analysis can be performed based on the precomputed polymorphic results to identify synonymous/nonsynonymous mutations and calculate silent (ds) and nonsilent (dn) mutation rates. As an analysis platform for IV research, IVDT continues to make enhancements to user interfaces, to improve infrastructures and functionality, and to add new applications. In order to identify the correlation between protein structure change and IV virulence, a key enhancement is to develop tools for displaying sequence variations in protein 3-D structures and to develop viewers for zooming in to a region of interest. More analytical tools, such as epitope prediction software and a primer design pipeline, will also be integrated.

## ACKNOWLEDGEMENTS

The work was sponsored by an institutional grant from the Beijing Institute of Genomics, Chinese Academy of Sciences. GFG is supported by National Basic Research Program 973 (Grant No. 2005CB523001) from Ministry of Science and Technology, China. We are indebted to all our collaborators, IV sample providers and other IV sequence producers for their contribution and kind support to the IVDB database. Funding to pay the Open Access publication charges for this article was provided by an institutional grant from the Beijing Institute of Genomics, CAS.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Horimoto,T. and Kawaoka,Y. (2005) Influenza: lessons from past pandemics, warnings from current incidents. *Nature Rev. Microbiol.*, **3**, 591–600.
2. Steinhauer,D.A. and Skehel,J.J. (2002) Genetics of influenza viruses. *Annu. Rev. Genet.*, **36**, 305–332.
3. Obenauer,J.C., Denson,J., Mehta,P.K., Su,X., Mukatira,S., Finkelstein,D.B., Xu,X., Wang,J., Ma,J., Fan,Y. *et al.* (2006) Large-scale sequence analysis of avian influenza isolates. *Science*, **311**, 1576–1580.
4. Ghedin,E., Sengamalay,N.A., Shumway,M., Zaborsky,J., Feldblyum,T., Subbu,V., Spiro,D.J., Sitz,J., Koo,H., Bolotov,P. *et al.* (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, **437**, 1162–1166.
5. Holmes,E.C., Ghedin,E., Miller,N., Taylor,J., Bao,Y., St George,K., Grenfell,B.T., Salzberg,S.L., Fraser,C.M., Lipman,D.J. *et al.* (2005) Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol.*, **3**, e300.
6. Xu,X., Subbarao, Cox,N.J. and 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**, 15–19.
7. Li,K.S., Guan,Y., Wang,J., Smith,G.J., Xu,K.M., Duan,L., Rahardjo,A.P., Puthavathana,P., Buranathai,C., Nguyen,T.D. *et al.* (2004) Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature*, **430**, 209–213.

8. Chen,H., Smith,G.J., Zhang,S.Y., Qin,K., Wang,J., Li,K.S., Webster,R.G., Peiris,J.S. and Guan,Y. (2005) Avian flu: H5N1 virus outbreak in migratory waterfowl. *Nature*, **436**, 191–192.
9. Kou,Z., Lei,F.M., Yu,J., Fan,Z.J., Yin,Z.H., Jia,C.X., Xiong,K.J., Sun,Y.H., Zhang,X.W., Wu,X.M. *et al*. (2005) New genotype of avian influenza H5N1 viruses isolated from tree sparrows in China. *J. Virol*., **79**, 15460–15466.
10. Liu,J., Xiao,H., Lei,F., Zhu,Q., Qin,K., Zhang,X.W., Zhang,X.L., Zhao,D., Wang,G., Feng,Y. *et al*. (2005) Highly pathogenic H5N1 influenza virus infection in migratory birds. *Science*, **309**, 1206.
11. Macken,C., Lu,H., Goodman,J. and Boykin,L. (2001) In Osterhaus,A. D. M. E., Cox,N. and Hampson,A. W. (eds), *Options for the Control of Influenza IV*. Elsevier Science, Amsterdam, pp. 103–106.
12. Kouranov,A., Xie,L., de la Cruz,J., Chen,L., Westbrook,J., Bourne,P.E. and Berman,H.M. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res*., **34**, D302–D305.
13. Webster,R.G., Bean,W.J., Gorman,O.T., Chambers,T.M. and Kawaoka,Y. (1992) Evolution and ecology of influenza A viruses. *Microbiol. Rev*., **56**, 152–179.
14. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
15. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.
16. Felsenstein,J. (1989) PHYLIP-Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.
17. Zhao,W., Wang,J., He,X., Huang,X., Jiao,Y., Dai,M., Wei,S., Fu,J., Chen,Y., Ren,X. *et al*. (2004) BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Res*., **32**, D377–D382.
18. Wang,J., He,X., Ruan,J., Dai,M., Chen,J., Zhang,Y., Hu,Y., Ye,C., Li,S., Cong,L. *et al*. (2005) ChickVD: a sequence variation database for the chicken genome. *Nucleic Acids Res*., **33**, D438–D441.
19. Wang,J., Xia,Q., He,X., Dai,M., Ruan,J., Chen,J., Yu,G., Yuan,H., Hu,Y., Li,R. *et al*. (2005) SilkDB: a knowledgebase for silkworm biology and genomics. *Nucleic Acids Res*., **33**, D399–D402.
20. Olsen,B., Munster,V.J., Wallensten,A., Waldenstrom,J., Osterhaus,A.D. and Fouchier,R.A. (2006) Global patterns of influenza a virus in wild birds. *Science*, **312**, 384–388.