Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W275–W279***
doi:10.1093/nar/gkm365

FluGenome: a web tool for genotyping influenza A virus

Guoqing Lu[1,2], Thaine Rowley[1,2], Rebecca Garten[3] and Ruben O. Donis[3,*]

[1]Department of Biology, [2]Department of Computer Science, University of Nebraska at Omaha, Omaha, NE and [3]Influenza Division, Centers for Disease Control and Prevention, Atlanta, GA

Received January 31, 2007; Revised April 3, 2007; Accepted April 25, 2007

ABSTRACT

Influenza A viruses are hosted by numerous avian and mammalian species, which have shaped their evolution into distinct lineages worldwide. The viral genome consists of eight RNA segments that are frequently exchanged between different viruses via a process known as genetic reassortment. A complete genotype nomenclature is essential to describe gene segment reassortment. Specialized bioinformatic tools to analyze reassortment are not available, which hampers progress in understanding its role in host range, virulence and transmissibility of influenza viruses. To meet this need, we have developed a nomenclature to name influenza A genotypes and implemented a web server, FluGenome (http://www.flugenome.org/), for the assignment of lineages and genotypes. FluGenome provides functions for the user to interrogate the database in different modalities and get detailed reports on lineages and genotypes. These features make FluGenome unique in its ability to automatically detect genotype differences attributable to reassortment events in influenza A virus evolution.

INTRODUCTION

Infections with influenza A viruses continue to be a public health problem, causing seasonal epidemics and sporadic but devastating pandemics. Each year in the US, influenza epidemics cause more than 200 000 hospitalizations and result in over 30 000 influenza-related deaths (1). Influenza pandemics are infrequent but they can result in high mortality. It is estimated that ~20–100 million people were killed worldwide by the 1918–1919 influenza pandemic (2–4). The current level of pandemic alert is at the highest level, phase 3, since the most recent pandemic of 1968 (5).

Influenza viruses belong to the family Orthomyxoviridae and are classified into three types, A, B and C based on the identity of major internal protein antigens (6). Influenza A and C viruses can infect multiple mammalian species, while influenza B virus is almost exclusively a human pathogen (7). Influenza A viruses cause the greatest morbidity and mortality in humans. Interestingly, the largest pool of influenza A viruses is maintained by horizontal spread in wild aquatic birds, in which the virus does not normally cause any disease (6,8). Food and companion animal populations such as poultry, swine, horses and dogs support sustained replication of certain lineages of influenza A, with minimal to lethal disease depending on the virulence of the strain (6). Influenza viruses have evolved in association with their various hosts in different continents for extended periods of time (9). This co-evolution has resulted in extensive genetic divergence among the extant viruses currently available for analysis.

Influenza A viruses are classified into subtypes on the basis of antigenic analysis of hemagglutinin (HA) and neuraminidase (NA) glycoproteins. So far, 16 HA subtypes and 9 NA subtypes have been found (10). In recent years, gene sequences have become available for a large number of viral strains creating a diverse pool of influenza A viruses from historical and current isolates collected in multiple geographic regions. Comparison of the deduced amino acid sequences of the HA and NA revealed an excellent agreement between the results of clustering viruses by the antigenic reactivity and sequence similarity. However, molecular genetic analysis allows a comprehensive analysis of the entire viral genome and is gaining popularity because it is more practical for most laboratories as a method for classification (11). Most importantly, study of the influenza genomic structure, namely genotyping, could reveal mechanisms of virus evolution, spread and disease pathogenesis.

The influenza A genome consists of eight negative-stranded RNA segments that encode at least 10 viral proteins (12). The viral genome evolves through accumulation of mutation by the viral RNA-dependent

*To whom correspondence should be addressed. Tel: 404 639 4968; Fax: 404 639 2350; Email: rdonis@cdc.gov

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

RNA polymerase which lacks proofreading ability and through reassortment of entire gene segments(13). Forces selecting viral variants such as the neutralizing antibody response of vertebrate hosts as well as species-related structural variation can also promote rapid evolution (14). Each of the segments can evolve at a different rate if they are subject to differential selective pressures and functional constraints (15–19). The segmented nature of the viral genome allows for segment exchange (termed reassortment) when two distinct viruses co-infect a cell and generate progeny with a mixed genome (20,21). Reassortment may theoretically yield 254 ($2^8 - 2$) different combinations of gene segments from two parent viruses.

A comprehensive influenza genotype database that can be searched using a web tool for the genotyping viruses is not available. Unlike HIV and HCV, the influenza A virus has a segmented genome, so eight separate phylogenies must be analyzed to establish a genotype.

We approached the problem of genotyping influenza A viruses by analyzing each gene segment independently, segregating gene segments into subtypes and subsequently into lineages. The genotype of an influenza A viral strain is the sequential aggregate of the eight assigned gene segment lineages. A nomenclature for influenza A viral genotypes will allow researchers to unequivocally describe influenza A viral genotypes to analyze, compare and communicate the molecular epidemiology of the virus. In this report, we define a nomenclature for influenza A viral genotypes and describe a web tool developed for genotyping influenza A viruses from genome sequences. Our tool facilitates identification of reassortment events between divergent lineages.

IMPLEMENTATION

Genotype nomenclature

Two nomenclature conventions are used routinely in influenza research: (i) the eight segments in the influenza A genome are numbered from 1 to 8 for PB2, PB1, PA, HA, NP, NA, M and NS, respectively; (ii) There are currently 16 alleles of the HA gene termed subtypes. Likewise, there are nine alleles for NA, and two alleles for non-structural (NS) proteins. Since influenza A viruses have an unusual genomic structure, we approached the genotyping problem by first analyzing each gene segment separately. According to the above conventions and considering that the evolutionary rate varies from segment to segment, we defined a genotype as a sequential combination of the lineages for each of the eight segments in a genome. A letter was assigned to each lineage of PB2, PB1, PA, NP and M, and a number followed by a letter was assigned to each lineage of HA, NA and NS with the number representing the subtype or allele. For example, [A,D,B,3A,A,2A,B,1A] is the genotype of a human seasonal subtype H3N2 virus with PB2 lineage A, PB1 lineage D, PA lineage B, HA subtype 3, lineage A and so on, following the convention for numbering of influenza genome segments. With this nomenclature, identifying genotypes and reassortment becomes an easy task accomplished by comparing the predicted genotype against all genomes that have been classified previously.

Lineage determination

Genomic sequences of all influenza A viruses with >75% of the full segment length were downloaded from NCBI Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html). Alignments were performed for each individual gene segment using the ClustalW program (22). The MEGA software was used to construct the phylogenetic trees with the neighbor-joining method and the HKY-85 model selected (23). The goal of our genotype method is to determine when a reassortment event with a gene segment from a non-traditional host or location has occurred. The lineages of each viral gene were carefully determined as detailed subsequently: (i) using the phylogenetic trees constructed, significant clusters (which were segregated by ~10% nucleotide difference by *p*-distance) were assigned lineages; (ii) bootstrap analysis was used on a smaller set of sequences with values >90% considered significant; (iii) the initial lineages were evaluated for nucleotide differences within and between other lineages and for strength of bootstrap support; (iv) approximately 10 sequences from each lineage were randomly selected for the maximum likelihood (ML) analysis for each gene segment, serotype (for HA, NA) or allele (for NS) on the MultiPhyl server (24). The lineage assignment of each influenza gene available in the public databases was uploaded into the Segment Table in the database as described subsequently.

Database

The FluGenome database contains three tables: Segment, Genome and Genotype. The Segment table contains information-related to sequences, including assigned lineage, strain name, segment, serotype, host, country, year, GenBank accession number, nucleotide sequence and sequence length. The Genome table contains the information for complete genomes, including assigned genotype and accession numbers of each gene segment. When more than one sequence was available for a gene segment, the longer of the two sequences was kept for the genome accession. Unique genotypes are stored in the Genotype table along with the total number of genomes that have that genotype. The Genotype table was created by querying the Genome table for distinct genotypes. Host categories were created to separate the genomes of each genotype, which include Human (Hu), Avian (Av), Swine (Sw), Equine (Eq), Canine (Ca) and Others (ONHM).

The FluGenome database is updated automatically every night. New sequences are downloaded from the NCBI Influenza Virus Resource (ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/) and added into the FluGenome database. The lineage information predicted for new sequences is used to update Segment, Genome and Genotype tables if necessary. For sequences already in the database, the script checks to see what information needs to be updated, and the sequences entries are flagged for further validation.

Web interface

The web interface and databases were implemented with the LAMP strategy. The server used Linux (L) for the operating system, along with Apache (A) as the web server. The genotyping database was built with the MySQL database management system (M). PHP and PERL (P) were used to code the two parts of the web tool: the back end program and the front end interface. JavaScript and HTML were used sparingly in the front end interface. A domain name, http://www.flugenome.org, was acquired to provide access to the database and the web tool.

Processing methods

The BLAST algorithm is used for sequence comparison, because of its advantages such as fast computation and accurate results in detecting local highly similar sequence regions. To overcome its inherent disadvantage (i.e. not a global alignment algorithm), we used a parameter called 'coverage' to detect gene-wide sequence similarity (25). The default thresholds for identifying lineages were set to be 95% identity and 95% coverage. The user can reset the thresholds to any allowable value. The top BLAST results for a user-submitted query sequence are sorted by identity and coverage, and the best result is used to assign a lineage to the query sequence. If a result from BLAST falls below the thresholds, the lineage will be flagged with an asterisk (*).

To determine the genotype of a complete or partial influenza virus genome, a script is executed that first establishes the lineage of each viral gene segment. The genotype will be created by the sequential incorporation of the lineages for each of the eight segments, arranged per convention as shown in Table 1. If a lineage does not meet the thresholds specified (95% default for both identity and coverage), the lineage will be assigned an asterisk (*) indicating the query sequence does not meet criteria and may be from a new lineage. If no BLAST results are found a blank lineage will be displayed. If all segments belong to known genotypes, the genotype of the query genomic sequence will be provided as output. The resulting genotype can be compared to previously identified genotypes in the Genotype database. This analysis can reveal reassortment events and host switching. If the genotype determined by FluGenome is not found in the Genotype database, the genome will be flagged as a virus with a potentially new genotype. Viral genotypes reported as new by FluGenome can simply result from identification of a gene from a novel phylogenetically defined lineage or the presence of genes from known lineages in novel combinations.

Table 1. The number of lineages derived and the number of sequences analyzed in each gene segment

Segment	No. of lineages	No. of sequences
PB2(1)	11	2955
PB1(2)	9	2822
PA(3)	11	2859
HA(4)[a]	78	6539
NP(5)	8	3252
NA(6)[a]	50	4013
MP(7)	7	3841
NS(8)[a]	10	3889

[a]HA and NA subtypes, and NS alleles are preserved.

FluGenome query options

The online tool presents two query options to the user; entering gene segment sequence(s) or genotype sequence(s) (Figure 1). The segment query 'Determine Individual Gene Segment Lineage' is used to identify the lineage of a viral gene segment of interest, for example PB2. In this case, the input FASTA file can contain one or many sequences, but all must correspond to the same gene segment. To analyze data sets from more than one gene simultaneously; e.g. both the PB1 and PB2, the user must first enter the number of different gene segments and then provide each sequence data set in a separate FASTA file.

The genotype query 'Determine Genotype' analyzes incomplete or complete genomes. Sequences from each genome must be in a separate FASTA file. Alternatively, the user can cut and paste sequences of one genome at a time. Multiple genomes can be analyzed simultaneously.

RESULTS

Genotype database

Nearly 30 000 sequences were collected from public databases and used for the lineage analyses, resulting in 184 lineages. The viral gene segments showed a wide range of diversity; HA was partitioned into 78 lineages whereas MP only into seven (Table 1). Mining the aforementioned sequences resulted in ~2300 complete genomes, which consists of 156 unique genotypes with 50 serotypes (http://www.flugenome.org/show_genotypes.php). Serotypes may comprise as many as 15, different genotypes;

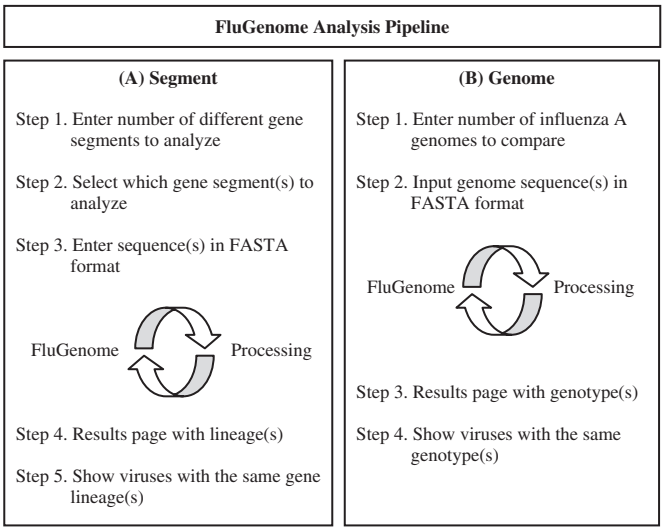

Figure 1. Schematic overview of analysis pipeline in FluGenome. (**A**) Segment analysis to determine the lineage of one or more gene segments from one or many different influenza A viruses. (**B**) Genome analysis to determine the genotype of one or more influenza A virus genomes.

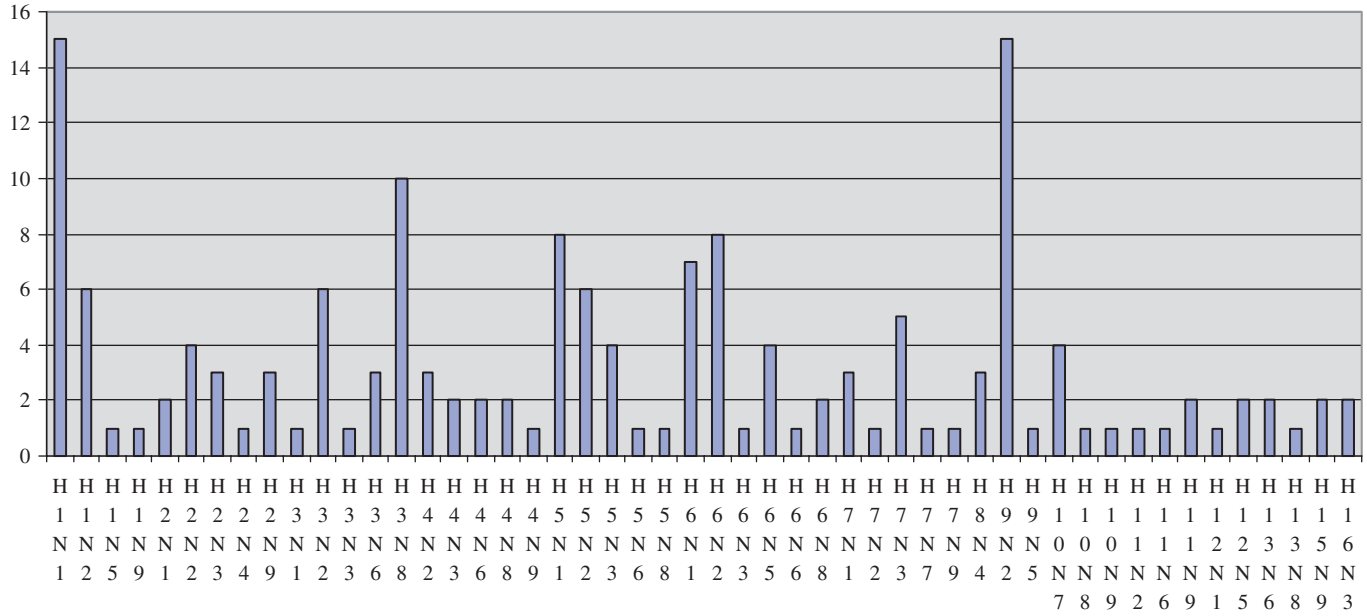

Figure 2. The number of genotypes observed in each serotype.

e.g. H1N1 and H9N2, whereas others just one (Figure 2). More than half of the complete genomes (1332) belong to the serotype H3N2 and have the genotype [A,D,B,3A,A,2A,B,1A].

Detecting reassortment

The FluGenome tool was designed to identify unique genotypes that arose by divergence as well as reassortment events between different circulating hosts. For example, in 1998, a serotype H3N2 virus was isolated from swine in North America, A/swine/Nebraska/209/1998 (26). Using FluGenome, this H3N2 virus had the genotype [C,D,E,3A,A,2A,A,1A] (Figure 3). It has been reported that this new genotype (termed 'triple reassortant') arose from reassortment events between human H3N2 viruses, swine H1N1 viruses and North American avian viruses of unknown serotype (27). The potential parent human H3N2 been circulating in humans since 1968 and has a genotype of [A,D,B,3A,A,2A,B,1A]. The triple reassortant virus acquired its HA, NA and PB1 gene segments from this seasonal human H3N2. The PB2 and PA genes arose from reassortment with avian viruses found in North America (with PB2 internal gene similar to A/mallard/ALB/126/1991 (CY005317) and PA gene similar to A/blue-winged teal/Alberta/141/1992 (CY004543)). The remaining three gene segments (NP, MP, NS) come from the classical H1N1 swine viruses whose genotype is [B,A,C,1A,A,1B,A,1A]. Although the NP and NS lineages are shared between the classical swine and human influenza viruses, the BLAST results show the closest matching isolates are swine in origin. Since 1998, this triple reassortant virus itself has undergone further reassortment with other swine and human influenza A viruses (28–31).

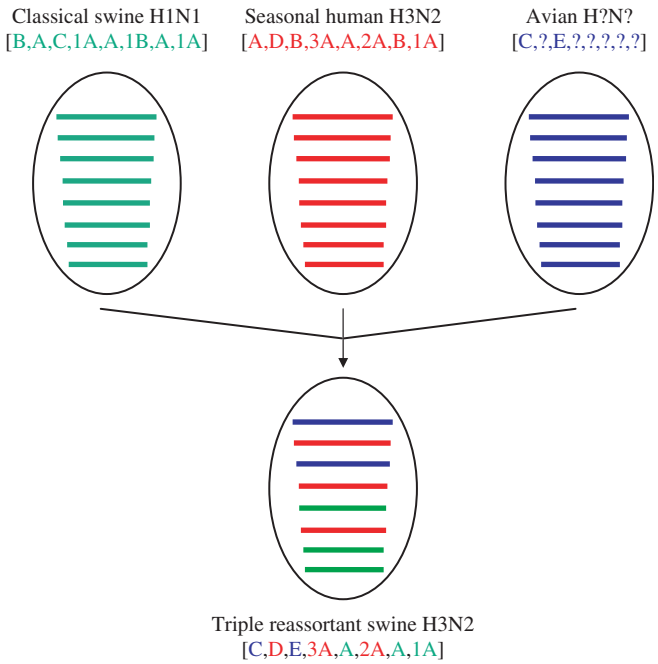

Figure 3. FluGenome detects the genesis of triple reassortant influenza A viruses isolated from swine.

CONCLUSION

We propose a nomenclature system for naming influenza A viral genotypes. This nomenclature was exploited to analyze ~2000 complete viral genomes (nearly full-length or full-length segment sequences), revealing 156 unique genotypes. The FluGenome web server implementation also includes facilities for analysis and sorting of lineages and genotypes which allow the user to explore the

evolutionary history of the viral strains. In particular, the FluGenome web server can provide genotype information that greatly facilitates the inference of genetic reassortment among influenza viruses.

ACKNOWLEDGEMENTS

The authors thank internal and outside users who tested FluGenome. We are grateful to Liying Jiang for her help with web programming at the early stage of this project. G.L. acknowledges the University of Nebraska at Omaha UCR grant for funding support. RG was supported in part by an Emerging Infectious Diseases (EID) Fellowship administered by the Association of Public Health Laboratories (APHL) and funded by the Centers for Disease Control and Prevention (CDC). Funding to pay the Open Access publication charges for this article was provided by CDC.

Conflict of interest statement. None declared.

REFERENCES

1. Thompson,W.W., Shay,D.K., Weintraub,E., Brammer,L., Cox,N., Anderson,L.J. and Fukuda,K. (2003) Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA*, **289**, 179–186.
2. Patterson,K.D. and Pyle,G.F. (1991) The geography and mortality of the 1918 influenza pandemic. *Bull. Hist. Med.*, **65**, 4–21.
3. Johnson,N.P. and Mueller,J. (2002) Updating the accounts: global mortality of the 1918–1920 "Spanish" influenza pandemic. *Bull. Hist. Med.*, **76**, 105–115.
4. Burnet,F.M. (1979) Portraits of viruses: influenza virus A. *Intervirology*, **11**, 201–214.
5. WHO. (2007), Vol. 2007.
6. 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.
7. Hay,A.J., Gregory,V., Douglas,A.R. and Lin,Y.P. (2001) The evolution of human influenza viruses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **356**, 1861–1870.
8. Widjaja,L., Krauss,S.L., Webby,R.J., Xie,T. and Webster,R.G. (2004) Matrix gene of influenza a viruses isolated from wild aquatic birds: ecology and emergence of influenza a viruses. *J. Virol.*, **78**, 8771–8779.
9. 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.
10. Fouchier,R.A., Rimmelzwaan,G.F., Kuiken,T. and Osterhaus,A.D. (2005) Newer respiratory virus infections: human metapneumovirus, avian influenza virus, and human coronaviruses. *Curr. Opin. Infect. Dis.*, **18**, 141–146.
11. Myers,R., Clark,C., Khan,A., Kellam,P. and Tedder,R. (2006) Genotyping Hepatitis B virus from whole- and sub-genomic fragments using position-specific scoring matrices in HBV STAR. *J. Gen. Virol.*, **87**, 1459–1464.
12. Lamb,R.A. and Choppin,P.W. (1983) The gene structure and replication of influenza virus. *Annu. Rev. Biochem.*, **52**, 467–506.
13. Drake,J.W. (1993) Rates of spontaneous mutation among RNA viruses. *Proc. Natl Acad. Sci. USA*, **90**, 4171–4175.
14. Matrosovich,M., Zhou,N., Kawaoka,Y. and Webster,R. (1999) The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distinguishable properties. *J. Virol.*, **73**, 1146–1155.
15. Altmuller,A., Fitch,W.M. and Scholtissek,C. (1989) Biological and genetic evolution of the nucleoprotein gene of human influenza A viruses. *J. Gen. Virol.*, **70**(Pt 8), 2111–2119.
16. Buonagurio,D.A., Nakada,S., Parvin,J.D., Krystal,M., Palese,P. and Fitch,W.M. (1986) Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene. *Science*, **232**, 980–982.
17. Ito,T., Gorman,O.T., Kawaoka,Y., Bean,W.J. and Webster,R.G. (1991) Evolutionary analysis of the influenza A virus M gene with comparison of the M1 and M2 proteins. *J. Virol.*, **65**, 5491–5498.
18. Gorman,O.T., Donis,R.O., Kawaoka,Y. and Webster,R.G. (1990) Evolution of influenza A virus PB2 genes: implications for evolution of the ribonucleoprotein complex and origin of human influenza A virus. *J. Virol.*, **64**, 4893–4902.
19. Okazaki,K., Kawaoka,Y. and Webster,R.G. (1989) Evolutionary pathways of the PA genes of influenza A viruses. *Virology*, **172**, 601–608.
20. Webster,R.G., Isachenko,V.A. and Carter,M. (1974) A new avian influenza virus from feral birds in the USSR: recombination in nature? *Bull. World Health Organ.*, **51**, 325–332.
21. Desselberger,U., Nakajima,K., Alfino,P., Pedersen,F.S., Haseltine,W.A., Hannoun,C. and Palese,P. (1978) Biochemical evidence that "new" influenza virus strains in nature may arise by recombination (reassortment). *Proc. Natl Acad. Sci. USA*, **75**, 3341–3345.
22. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
23. Kumar,S., Tamura,K. and Nei,M. (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.*, **5**, 150–163.
24. Keane,T.M., Naughton,T.J., Travers,S.A., McInerney,J.O. and McCormack,G.P. (2005) DPRml: distributed phylogeny reconstruction by maximum likelihood. *Bioinformatics*, **21**, 969–974.
25. Lu,G., Jiang,L., Helikar,R.M., Rowley,T.W., Zhang,L., Chen,X. and Moriyama,E.N. (2006) GenomeBlast: a web tool for small genome comparison. *BMC Bioinformatics*, **7**(Suppl. 4), S18.
26. Karasin,A.I., Schutten,M.M., Cooper,L.A., Smith,C.B., Subbarao,K., Anderson,G.A., Carman,S. and Olsen,C.W. (2000) Genetic characterization of H3N2 influenza viruses isolated from pigs in North America, 1977–1999: evidence for wholly human and reassortant virus genotypes. *Virus Res.*, **68**, 71–85.
27. Zhou,N.N., Senne,D.A., Landgraf,J.S., Swenson,S.L., Erickson,G., Rossow,K., Liu,L., Yoon,K.J., Krauss,S. *et al.* (2000) Emergence of H3N2 reassortant influenza A viruses in North American pigs. *Vet. Microbiol.*, **74**, 47–58.
28. Karasin,A.I., Landgraf,J., Swenson,S., Erickson,G., Goyal,S., Woodruff,M., Scherba,G., Anderson,G. and Olsen,C.W. (2002) Genetic characterization of H1N2 influenza A viruses isolated from pigs throughout the United States. *J. Clin. Microbiol.*, **40**, 1073–1079.
29. Webby,R.J., Rossow,K., Erickson,G., Sims,Y. and Webster,R. (2004) Multiple lineages of antigenically and genetically diverse influenza A virus co-circulate in the United States swine population. *Virus Res.*, **103**, 67–73.
30. Ma,W., Gramer,M., Rossow,K. and Yoon,K.J. (2006) Isolation and genetic characterization of new reassortant H3N1 swine influenza virus from pigs in the midwestern United States. *J. Virol.*, **80**, 5092–5096.
31. Karasin,A.I., West,K., Carman,S. and Olsen,C.W. (2004) Characterization of avian H3N3 and H1N1 influenza A viruses isolated from pigs in Canada. *J. Clin. Microbiol.*, **42**, 4349–4354.