Epidemiol. Infect. (2006), **134**, 514–520. © 2005 Cambridge University Press
doi:10.1017/S0950268805005686 Printed in the United Kingdom

Detailed analysis of the genetic evolution of influenza virus during the course of an epidemic

A. LAVENU[1,2], M. LERUEZ-VILLE[3], M.-L. CHAIX[3], P.-Y. BOELLE[1,2,6], S. ROGEZ[4], F. FREYMUTH[5], A. HAY[7], C. ROUZIOUX[3] AND F. CARRAT[1,2,6]*

[1] *INSERM U707, Paris, France*
[2] *Université Paris 6, Paris, France*
[3] *Laboratoire de Virologie, EA 3620 Université René Descartes. CHU Necker Enfants-Malades, Paris, France*
[4] *Laboratoire de Bactériologie, Virologie et Hygiène, CHU Dupuytren, Limoges, France*
[5] *Laboratoire de Virologie humaine et moléculaire, CHU Côte de Nâcre, Caen, France*
[6] *Assistance Publique des Hôpitaux de Paris, Hôpital Saint Antoine, Paris, France*
[7] *Virology Division, National Institute for Medical Research, The Ridgeway, London, UK*

(*Accepted 13 October 2005, first published online 29 November 2005*)

SUMMARY

The genetic variability of influenza virus is usually studied with sequences selected over numerous years and countries, and rarely within a single season. Here we examined the viral evolution and the correlation between genetic and clinical features during an epidemic. From a French prospective household-based study in 1999–2000, 99 infected patients were randomly selected. The HA1 genomic domain was sequenced. Phylogenetic analysis showed the existence of two groups of A/H3N2 viruses. We found no distinct pattern of genomic evolution within either group according to time. A spatial correlation with the nucleotide distances was shown. The average nucleotide diversity was $3{\cdot}4 \times 10^{-3}$ nucleotides per site, and did not differ between the groups. A lower number of segregating sites was observed in patients who experienced influenza-like symptoms during the previous epidemic. These results suggest that the influenza virus undergoes regular HA1 nucleotide changes, but without clonal expansion of mutant strains within a single epidemic.

INTRODUCTION

Influenza virus shows particularly rapid genetic evolution, notably in the haemagglutinin (HA) and neuraminidase (NA) genes. Evolution of the HA1 domain of the HA gene has been studied worldwide for more than 10 years [1–6]. These studies show that the genetic evolution follows a single lineage for the A/H3 subtype and more than one lineage for the A/H1 subtype and type B viruses. Rates of nucleotide changes range between $1{\cdot}3\times10^{-3}$ and $3{\cdot}7\times10^{-3}$ per site per year.

Numerous published data are available on antigenic or genetic evolution over shorter periods, for example 1–5 winter seasons [7–14]. They provide useful results on the genetic and antigenic properties of circulating strains, identify yearly changes and help for a better selection of appropriate vaccine strains.

However, in all these studies the sampling method used to produce datasets has a number of problems that might affect a rate estimate of nucleotide variation. These include the low number of samples per winter season, temporal or geographic bias, lack of randomness in selection. In addition many strains

* Author for correspondence: Dr F. Carrat, INSERM U707, Faculté de Medecine St Antoine, 27 rue de Chaligny, 75012 Paris, France.
(Email: carrat@u707.jussieu.fr)

selected for sequencing are associated with atypical clinical features, and most published sequences correspond to highly variant viruses [15].

Knowledge of short-term evolution is nonetheless important for understanding the general evolutionary dynamics of influenza viruses, and might help to identify the most likely candidate strain for next season's epidemic. This would have obvious implications for year-on-year vaccine development.

Here we studied the genetic evolution of influenza virus during a single epidemic in France, using rigorous strain selection criteria. We also examined the possible correlation between genetic variability, pathogenicity and transmissibility.

METHODS

Isolates

This study was part of a large prospective household contact study done during the French influenza epidemic of winter 1999–2000 [16]. Briefly, 161 general practitioners from the French Sentinel Network participated in the study. A household was enrolled when a member visited a general practitioner with the following characteristics: fever ($\geqslant 38\ ^{\circ}C$) starting <48 h previously; respiratory signs; at least one other person present in the household; first household case; ambulatory treatment; and informed consent. These patients were considered as the household index cases. A nasal swab was taken for centralized virological studies (direct immunofluorescence, cell culture and PCR). All household members, including the index patient, were monitored for 15 days (symptoms, health-care use). In total, 946 index patients were enrolled in the main study; 510 index cases with laboratory-confirmed A/H3N2 influenza were identified; and 395 patients (77%) completed the follow-up period. Ninety-nine isolates were selected for sequencing, by means of a randomized sampling protocol based on time (beginning, middle and end of the epidemic) and space (21 administrative regions).

PCR amplification and sequencing

Genomic RNA was extracted from 140 μl of nasal swab fluid with the QIA amp RNA mini kit (Qiagen, Courtaboeuf, France) following the manufacturer's recommendations.

The HA1 genomic domain was amplified and sequenced. Complementary DNA synthesis and first-round PCR amplification were performed with forward primer 5′-CTATCATTGCTTTGAGC-3′ and reverse primer 5′-ATGGCTGCTTGAGTGCTT-3′, as described by Fitch *et al*. [5]. This step used the BM-Titan one-tube RT–PCR procedure (Roche Diagnostic, Meylan, France). Nested PCR was then performed with forward primer 5′-GAGCTACA-TTTTTATGTCTGGT-3′ and reverse primer 5′-GTGCTTTTAAGATCTGCTGCT-3′. Final PCR products (1135 nucleotides long) were visualized by agarose gel electrophoresis and purified with the QIAquick PCR purification kit (Qiagen). After purification, PCR products were sequenced using the fluorescent dideoxy-terminator method (Big Dye Terminator kit, Applied Biosystems, PerkinElmer, Foster City, CA, USA) on an Applied Biosystems 377 automated DNA sequencer (ABI/PE, Applied Biosystems). The following internal primers were used for sequencing: forward 5′-AGTCACAGTCTCTACCA-A-3′, reverse 5′-GGTGCAACCAATTCAATCTA-3′. Both strands were sequenced. Sequences were aligned with Sequence Navigator® software (Applied Biosystems). The nucleotide sequences determined in this study are available on the GenBank and Los Alamos National Laboratory Influenza Sequence databases [17] (accession numbers AY633996–AY634049).

Analysis

The nucleotide sequences of the HA1 genomic domain (987 bp) were translated into amino-acid sequences (329 amino acids) in order to distinguish synonymous from non-synonymous changes [18, 19]. Among the 329 amino acids, 130 lie within or close to the five main epitopes of the haemagglutinin protein, labelled A–E [20]. A consensus sequence was constructed from the most frequent nucleotides at each site.

Two estimates of DNA sequence variation were used. Nucleotide diversity is the average proportion of different nucleotides between all pairs of sequences. The standardized number (θ) of segregating sites (S) is the number of sites with at least two different nucleotides, standardized by a factor proportional to the number of sequences studied [21]. The nucleotide distance is the number of nucleotides differing between two sequences; spatial distance is the Euclidean distance between the sources of two nasal swabs (in km); and temporal distance is the number of days separating two swabs.

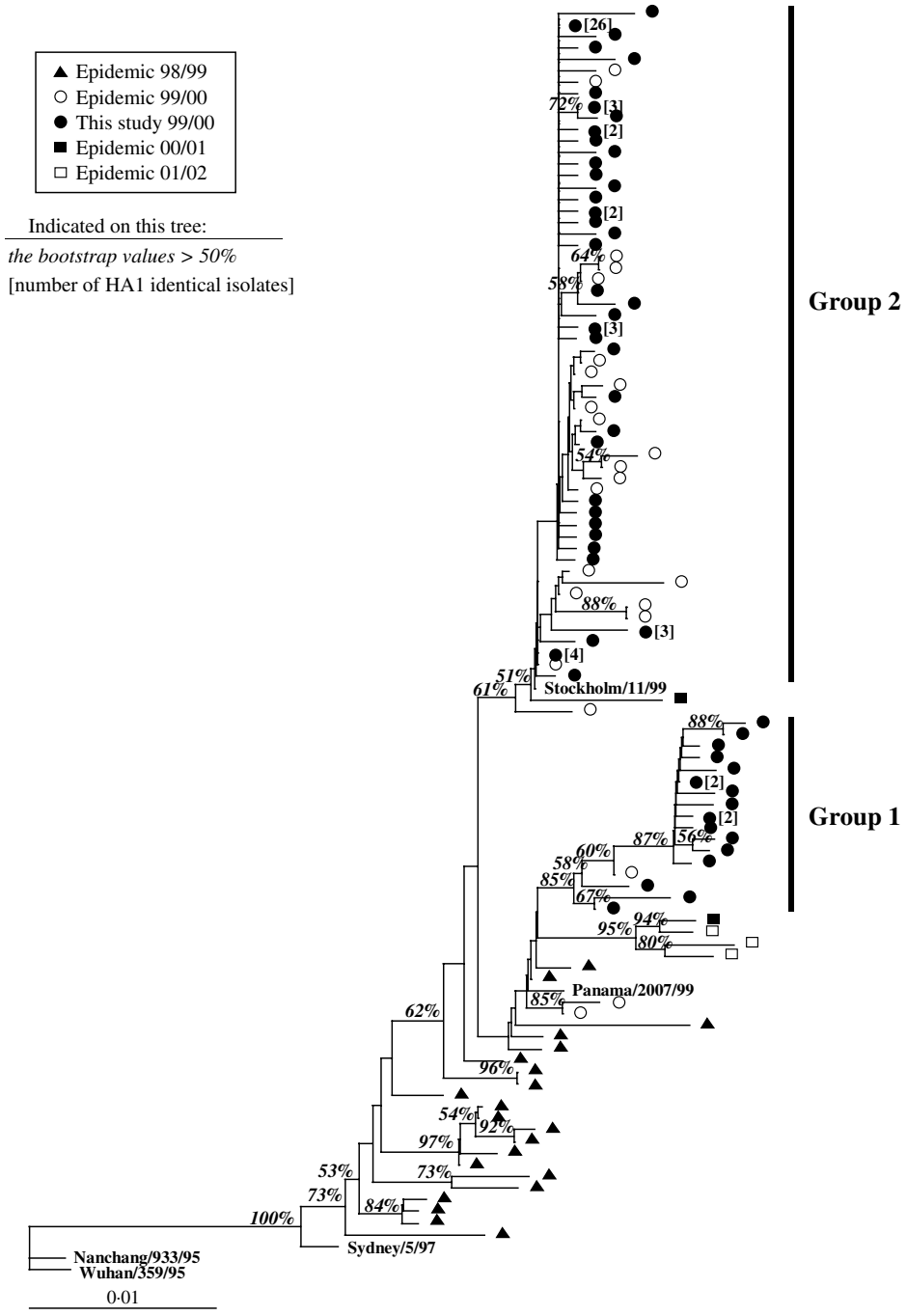

Fig. 1. Phylogenetic analysis of the evolution of the HA1 domain of influenza virus A(H3N2) during winter 1999–2000. Strains representative of H3N2 clusters identified between 1995 and 2000 [1], and European strains isolated between 1999 and 2002, were also integrated. The tree rooted on Wuhan/359/95 was generated by using the Neighbor-Joining method. The lengths of the horizontal lines are proportional to the number of nucleotide substitutions. The trees were bootstrapped 100 times. The number of identical isolates on each branch representing each sequence is noted when >1.

Viral evolution was first studied phylogenetically. Multiple sequence alignments were performed with ClustalW software version 1.7 [22]. The nucleotide sequences for A/Wuhan/359/95, A/Nanchang/933/95, A/Sydney/5/97, A/Panama/2007/99 and A/Stockholm/11/99 (representative members from four clusters of H3N2 isolates obtained between 1995 and 2000 [1]) were obtained from the public Los Alamos National Laboratory Influenza Sequence Database [17] under accession numbers AF008722, AF008725, AJ311466, ISDNCDA001 and ISDNSWA001. The 50 European H3N2 strains isolated between 1999 and 2002 were

https://doi.org/10.1017/S0950268805005686 Published online by Cambridge University Press

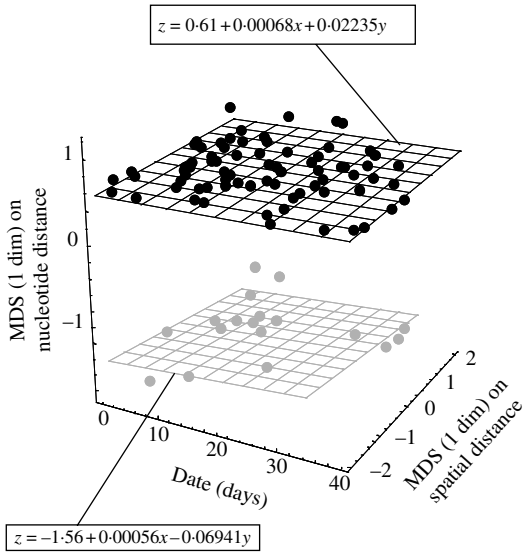

Fig. 2. Distribution of influenza virus isolates according to the nucleotide distance and the spatial distance [each distance is represented by multidimentional scaling (MDS) with one dimension, and is referred to as the nucleotide MDS distance and spatial MDS distance respectively] and according to the date of isolation. Group 1 is in grey and Group 2 in black. Plans estimated for each group representing the nucleotide MDS distance according to the date and the spatial MDS distance were drawn.

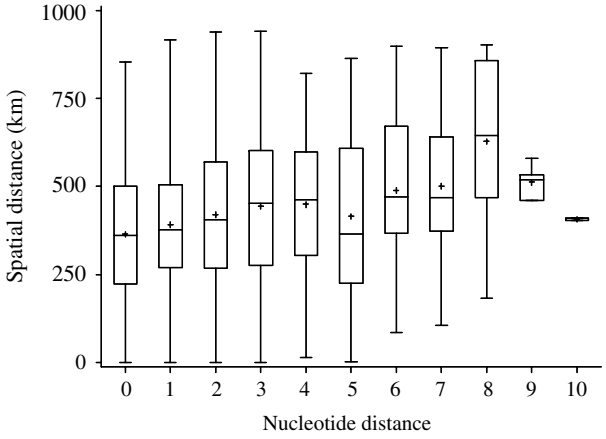

Fig. 3. Distribution of the pairwise Group 2 isolates according to the spatial distance for each nucleotide distance (Box plots of 2701 points).

obtained from the same database under accession numbers AF357931–33, AF357944–54, AF357966, AF357968–69, AF442460, AF442478, AF442481, AY661021–23, AY661026, AY661031, ISDN13294, ISDN13326, ISDNOS0012, ISDNOS0016–19, ISDNOS0022, ISDNOS99, ISDNSW0004–10, ISDNSW0014–15, ISDNSW0018–19, ISDNSWA001–03, ISDNSWA010, ISDNSWA013. The phylogenetic tree rooted on A/Wuhan/359/95 was built from the nucleotide distances by using the Neighbor-Joining method (NEIGHBOR program of PHYLIP software version 3.5c) [23]. The tree was drawn with TreeView version 1.6.6. The bootstrap values were obtained by using 100 replicates of the tree [24].

We then studied simultaneously how the nucleotide distances correlated with space and time distances by using the multidimensional scaling (MDS) method. This method has already been used to analyse influenza virus antigenic variability based on haemagglutination-inhibition [25]. Its main advantage is to map the distances between all strains in a one-dimensional space [26].

The Kruskal–Wallis method was used to compare continuous variables and Fisher's exact test was used for categorical variables. Mantel's test was used to measure correlations between distance matrixes. The Log-rank test was used to analyse the duration of disease. The standardized number of segregating sites was compared between groups of strains by taking into account the latent dating of a common ancestor in the calculation of covariance [27]. Extreme time-lags between the strains and the common ancestor (0 and infinity) were used to obtain the minimal and maximal P values. For clarity, the maximal P value is given if both are $<0{\cdot}05$, the minimum if both are $>0{\cdot}05$, and none if the minimal is $<0{\cdot}05$ and the maximal $>0{\cdot}05$.

RESULTS

Nucleotide sequence analysis

Ninety-two isolates were successfully amplified. Amplification failed in the other cases because of poor sample quality; these samples did not differ from the amplified samples in terms of the place or date of isolation, or clinical severity.

Two distinct groups of isolates were revealed by phylogenic analysis (Fig. 1). Group 1 was composed of 18 isolates corresponding to 16 strains. Group 2 was composed of 74 isolates corresponding to 38 strains. Group 1 was close to the strain A/Panama/2007/99 and Group 2 to A/Stockholm/11/99. In Group 2, 26 isolates were identical. No group ancestor was identified among the isolates. Three strains from Group 1 were closer to strains isolated during subsequent epidemics than were the remainder of Group 1 strains. However, these three strains were also closer to strains isolated during the previous epidemic.

https://doi.org/10.1017/S0950268805005686 Published online by Cambridge University Press

Table 1. *Nucleotide differences between the consensus sequences of the two groups*

Nucleotide position	14	48	99	2611	27565	26	431	471	483	633	774	799	811	817
Group 1	G	G	A	C	A	G	A	G	C	G	C	G	G	T
Group 2	T	A	C	T	C	A	T	A	T	A	T	A	A	C
Non-synonymous	√	—	√	—	√	—	√	—	—	—	—	√	√	√
Epitopes	—	—	—	E	E	A	A	B	—	—	—	—	—	C

Group 2 consensus strain is also the consensus strain of the total sample.

Table 2. *Genetic variation among Group 2 haemagglutinin of patients exposed and unexposed to different clinical and epidemiological factors. θ is the standardized number of segregating sites* (*P values were computed according to* [*27*]*, as described in the Methods section*)

Clinical and epidemiological factors in Group 2 (74 patients)	Yes		No		
	n	θ	*n*	θ	*P*
Age > 20 yr*	52	7·33	22	6·26	> 0·423
Age > 65 yr*	8	3·47	66	10·00	n.d.
Symptom score > 0·7	34	5·87	39	7·05	> 0·277
Duration of illness in days > 8	29	4·67	44	7·56	n.d.
Influenza in the preceding season	9	0·74	62	10·30	< 0·005
Vaccination the same season	7	4·90	65	9·24	> 0·114
Size of household > 3	35	7·14	39	6·82	> 0·780
Secondary cases	35	6·28	39	7·04	> 0·490

n.d., Not defined (if the minimal *P* value is < 0·05 and the maximal > 0·05, *P* = 'not defined').
Threshold values in the factor labels are the median in Group 2, except for age (*) which differentiates young and old persons from others. Secondary cases are defined as members of the household having influenza symptoms within 5 days after the first case [32].

Nucleotide distance analysis in time and space

The MDS method identified two distinct groups in terms of nucleotide distances ($P < 0{\cdot}0001$, Fig. 2). The two groups were strictly identical to those obtained by phylogenic analysis. The temporal and spatial distances did not differ between the groups ($r = -0{\cdot}031$, $P = 0{\cdot}77$ and $r = 0{\cdot}009$, $P = 0{\cdot}39$ respectively). No correlation was found between the nucleotide distance and the temporal distance (Group 1: $r = -0{\cdot}136$, $P = 0{\cdot}83$; Group 2: $r = -0{\cdot}008$, $P = 0{\cdot}51$), however, the nucleotide distance and the spatial distance were correlated within Group 2 (Group 1: $r = 0{\cdot}068$, $P = 0{\cdot}29$; Group 2: $r = 0{\cdot}178$, $P = 0{\cdot}001$) (Fig. 3).

Nucleotide diversity

Nucleotide diversity was $4{\cdot}6 \times 10^{-3}$ nucleotides per site in Group 1 and $2{\cdot}2 \times 10^{-3}$ nucleotides per site in Group 2. The standardized number of segregating sites was 8·14 and 10·21 respectively, and did not differ between the groups ($P > 0{\cdot}29$). The consensus sequences of the two groups differed by 14 sites (Table 1). Seven of these sites were non-synonymous, of which three were located within or near one of the main epitopes of the haemagglutinin protein. The proportion of non-synonymous differences located within or near an epitope did not differ from that expected by chance (0·43%, $P = 1{\cdot}00$).

Clinical impact and transmissibility

The two groups did not differ in terms of clinical or epidemiological factors (not shown). Owing to the small number of isolates in Group 1, the comparison of nucleotide diversity according to the presence or absence of clinical or epidemiological factors was limited to Group 2. We found a lower standardized number of segregating sites in strains infecting patients who had had 'flu-like syndromes during the previous season (Table 2).

DISCUSSION

Two groups of H3N2 influenza strains, one close to A/Stockholm/11/99 and the other to A/Panama/2007/99

https://doi.org/10.1017/S0950268805005686 Published online by Cambridge University Press

were clearly identified by both phylogenetic analysis and MDS. These results are in keeping with the analysis of worldwide surveillance data for 1999–2000, which showed expansion of the A/Stockholm/11/99 cluster and regression of the clusters represented by A/Panama/2007/99 and A/Sydney/5/97 [1]. The fact that some isolates belonging to the two groups were found during the same period and at the same location might be interpreted as reflecting two 'simultaneous' epidemics, which is not a rare phenomenon [28, 29]. We found that, within the group (Group 2) including the highest number of isolates, nucleotide distances between strains and spatial distances were correlated. This correlation may be explained by the local spread of strains which, although belonging to the same group, exhibit slightly different sequences according to their geographic origin. By contrast, no correlation was observed between nucleotide and temporal distances suggesting no clear evolutionary pattern during the epidemic.

Several results warrant discussion. First, the lack of data on the rate of evolution of influenza virus during a single epidemic made it difficult for us to compute the number of isolates needed to demonstrate a link between nucleotide distances and time. With the number of isolates studied, the statistical power may have been low. However, we obtained relatively low correlations and a marked increase in the sample size would have been unlikely to improve the pertinence of the results. Second, we cannot exclude the possibility that different results would have been obtained in other years, but it is known that the evolutionary rate of A/H3N2 strains is relatively constant year on year [5]. Third, we only sequenced the HA1 genomic domain; although most studies of the genetic diversity of influenza virus are restricted to HA1, the evolution of HA and NA may be different [8].

Average nucleotide diversity was $3{\cdot}4 \times 10^{-3}$ nucleotides per site, while the average standardized number of segregating sites was 9·17 and did not differ between the two groups of strains. The standardized number of segregating sites was lower in strains infecting patients who had had 'flu-like syndromes during the previous season. Indeed, the proportions of strains identical to the consensus sequence was higher in these patients. One possible explanation is that frail patients might be more susceptible to strains showing little genetic change than are patients in good health.

The observed lack of correlation between the viral group and clinical manifestations may be explained by a lack of statistical power, the exclusion of asymptomatic infections, or by a real lack of difference in pathogenicity. This latter possibility is supported by a recent study in which no difference in disease severity was observed between children infected with different A/H3N2 strains [30].

This study shows that the influenza virus undergoes regular HA1 nucleotide changes, but without clonal expansion of mutant strains. Since there is Darwinian and continuing forward selection of influenza virus variants when a longer time period (years rather than weeks) is studied [31], we concluded that the evolution of strains year on year may be due mainly to stronger positive selection during non-epidemic periods, or to worldwide spread with regular, minor changes. For this reason, no particular strain isolated during an epidemic can be considered the closest to strains of the subsequent epidemic.

ACKNOWLEDGEMENTS

We acknowledge the financial support of GlaxoSmithKline Inc. for collection of data and viruses.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Plotkin JB, Dushoff J, Levin SA.** Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proceedings of the National Academy of Sciences USA* 2002; **99**: 6263–6268.
2. **Levin SA, Dushoff J, Plotkin JB.** Evolution and persistence of influenza A and other diseases. *Mathematical Biosciences* 2004; **188**: 17–28.
3. **Ferguson NM, Galvani AP, Bush RM.** Ecological and immunological determinants of influenza evolution. *Nature* 2003; **422**: 428–433.
4. **Bush RM, *et al.*** Predicting the evolution of human influenza A [see comments]. *Science* 1999; **286**: 1921–1925.
5. **Fitch WM, *et al.*** Long term trends in the evolution of H(3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences USA* 1997; **94**: 7712–7718.
6. **Hay AJ, *et al.*** The evolution of human influenza viruses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 2001; **356**: 1861–1870.
7. **Nakagawa N, *et al.*** Heterogeneity of influenza B virus strains in one epidemic season differentiated by monoclonal antibodies and nucleotide sequences. *Journal of Clinical Microbiology* 2000; **38**: 3467–3469.
8. **Abed Y, *et al.*** Divergent evolution of hemagglutinin and neuraminidase genes in recent influenza A:H3N2

viruses isolated in Canada. *Journal of Medical Virology* 2002; **67**: 589–595.
9. **Pyhala R, *et al.*** Influence of antigenic drift on the intensity of influenza outbreaks: upper respiratory tract infections of military conscripts in Finland. *Journal of Medical Virology* 2004; **72**: 275–280.
10. **Pyhala R, *et al.*** Phylogenetic and antigenic analysis of influenza A(H3N2) viruses isolated from conscripts receiving influenza vaccine prior to the epidemic season of 1998/9. *Epidemiology and Infection* 2002; **129**: 347–353.
11. **Ellis JS, Zambon MC.** Molecular analysis of an outbreak of influenza in the United Kingdom. *European Journal of Epidemiology* 1997; **13**: 369–372.
12. **Ellis JS, Chakraverty P, Clewley JP.** Genetic and antigenic variation in the haemagglutinin of recently circulating human influenza A (H3N2) viruses in the United Kingdom. *Archives of Virology* 1995; **140**: 1889–1904.
13. **Pontoriero AV, *et al.*** Antigenic and genomic relation between human influenza A (H3N2) viruses circulating in Argentina during 1998 and the H3N2 vaccine component. *Revista Panamericana de Salud Publica* 2001; **9**: 246–253.
14. **Pontoriero AV, *et al.*** Antigenic and genomic relation between human influenza viruses that circulated in Argentina in the period 1995–1999 and the corresponding vaccine components. *Journal of Clinical Virology* 2003; **28**: 130–140.
15. **Bush RM, *et al.*** Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *Proceedings of the National Academy of Sciences USA* 2000; **97**: 6974–6980.
16. **Carrat F, *et al.*** Influenza burden of illness: estimates from a national prospective survey of household contacts in France. *Archives of Internal Medicine* 2002; **162**: 1842–1848.
17. **Macken C, *et al.*** The value of a database in surveillance and vaccine selection. In: Osterhaus ADME, Cox N, Hampson AW, eds. *Options for the Control of Influenza IV*. Amsterdam: Elsevier Science, 2001: pp. 103–106.
18. **Yang Z, Bielawski JP.** Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution* 2000; **15**: 496–503.
19. **Bush RM, *et al.*** Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Molecular Biology and Evolution* 1999; **16**: 1457–1465.
20. **Plotkin JB, Dushoff J.** Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proceedings of the National Academy of Sciences USA* 2003; **100**: 7152–7157.
21. **Tajima F.** Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989; **123**: 585–595.
22. **Thompson JD, Higgins DG, Gibson TJ.** CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 1994; **22**: 4673–4680.
23. **Felsenstein J.** PHYLIP: Phylogeny inference package, version 3.572. *Cladistics* 1989; **5**: 164–166.
24. **Zharkikh A, Li WH.** Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Molecular Biology and Evolution* 1992; **9**: 1119–1147.
25. **Smith DJ, *et al.*** Mapping the antigenic and genetic evolution of influenza virus. *Science* 2004; **305**: 371–376.
26. **Hardle W, Simar L.** Metric multidimensional scaling. In: *Applied Multivariate Statistical Analysis*. Berlin, New York: Springer Verlag, 2003: pp. 379–382.
27. **Innan H, Tajima F.** A statistical test for the difference in the amounts of DNA variation between two populations. *Genetical Research* 2002; **80**: 15–25.
28. **Simonsen L, *et al.*** The impact of influenza epidemics on mortality: introducing a severity index. *American Journal of Public Health* 1997; **87**: 1944–1950.
29. **Lavenu A, Valleron AJ, Carrat F.** Exploring cross-protection between influenza strains by an epidemiological model. *Virus Research* 2004; **103**: 101–105.
30. **O'Donnell FT, *et al.*** Epidemiology and molecular characterization of co-circulating influenza A/H3N2 virus variants in children: Houston, Texas, 1997–8. *Epidemiology and Infection* 2003; **130**: 521–531.
31. **Fitch WM, *et al.*** Positive Darwinian evolution in human influenza A viruses. *Proceedings of the National Academy of Sciences USA* 1991; **88**: 4270–4274.
32. **Gubareva LV, Novikov DV, Hayden FG.** Assessment of hemagglutinin sequence heterogeneity during influenza virus transmission in families. *Journal of Infectious Diseases* 2002; **186**: 1575–1581.