JOURNAL OF VIROLOGY, Sept. 2008, p. 8947–8950 Vol. 82, No. 17
0022-538X/08/$08.00+0 doi:10.1128/JVI.00101-08
Copyright © 2008, American Society for Microbiology. All Rights Reserved.

Anomalies in the Influenza Virus Genome Database: New Biology or Laboratory Errors?▽†

Michael Krasnitz,*‡ Arnold J. Levine, and Raul Rabadan*‡

Institute for Advanced Study, Einstein Dr., Princeton, New Jersey 08540

Received 15 January 2008/Accepted 6 June 2008

A search of the influenza virus genome database reveals anomalies associated with a nonnegligible number of submitted sequences. There are many pairs of viral segments that are very close to each other in nucleotide sequence but relatively far apart in reported time of isolation, resulting in an abnormally low evolutionary rate. Also, some sequences show clear evidence of apparent homologous recombination, a process normally assumed to be extremely rare or nonexistent in this virus. These findings may point to surprising new biology but are perhaps more readily explained by stock contamination or other errors in the sequencing laboratories.

In the last few years, an extraordinary amount of influenza virus genomic sequence has been submitted to publicly available databases (see, e.g., http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html, http://www.flu.lanl.gov, and http://influenza.genomics.org.cn). For instance, there are now over 3,300 full genome sets in the NCBI's rapidly growing Influenza Virus Resource. To our knowledge, no systematic attempt has been made to assess the quality of sequence data in this and similar collections. Our observations show that a fraction of the sequences in the database exhibit anomalous properties that point to either radically new biology or, more likely, problems with the data. As a first example, we consider the rate of nucleotide substitution in the influenza A virus. This rate has been previously estimated at 0.001 to 0.007 per nucleotide per year. (There have been many studies analyzing influenza virus evolutionary rates in different segments and different hosts; see, among others, references 6, 8, 9, 10, 11, 14, and 16.) Using the most conservative (lowest) estimate, we still find many pairs of virus segments that are far closer to each other in nucleotide space than would randomly occur in a Poisson process with this evolutionary rate, given the difference in time of isolation. Such sequences appear to be effectively "frozen in time." For instance, the PB2 segments of isolates A/duck/Taiwan/0526/1972(H6N1) and A/chicken/Taiwan/G23/87(H6N1) differ in only 1 nucleotide position out of 2,283 aligned nucleotides, whereas the expected number of differences, at 0.0015 substitution per nucleotide per year, would be ~48 for 15 years. For a null Poisson process, this gives an extremely low P value of 6.6×10^{-20}. Note also that 15 years is actually a lower bound on the true evolutionary time between these two segments, since their latest common ancestor is likely to predate both; this makes their virtual identity even more improbable. To visualize this anomaly, consider the plot in Fig. 1.

One can see that the great majority of segment pairs lie on or above the dashed line representing a rough estimate of the expected number of nucleotide differences given the distance in time between the isolates; most are above the line since the true evolutionary time (the combined distance to the latest common ancestor) is generally greater than the naive distance in years. However, a number of points lie significantly below the fit curve, with corresponding extremely low P values. These represent viruses that appear to be "frozen in time." We performed a systematic search for such "frozen" sequences; for the results, see Appendix S1 in the supplemental material. We found about 60 isolates which show strong evidence of an anomalously low apparent evolutionary rate, with highly significant Bonferroni-corrected P values. Most of these are viruses from avian and swine hosts, many of them H5N1 isolates submitted from Asia in recent years. The phenomenon of "frozen evolution" occurs at roughly the same rate in all influenza virus segments, allowing for greater statistical power to detect it in the longer segments. Often, though not always, multiple segments of the same isolate appear to have evolved at a very low rate with respect to an ancestor virus. These results are insensitive to the exact rate estimates and methods used to compute the P values. Each of the anomalous sequences is so close to some other sequence in the database compared to their distance in time that any model would rule out the possibility of a random fluctuation; there is clearly something extraordinary about these sequences. As discussed below, we omitted the human H1N1 viruses corresponding to the mysterious reemergence of H1N1 in 1977, as well as a few other suspect isolates previously reported in the literature (1, 2, 3, 7, 23) (S. H. Seo, J. A. Kim, and S. K. Jo, unpublished data, 2004).

Another anomaly present in the influenza virus database relates to homologous recombination. This mechanism is generally believed to be extremely rare or nonexistent in the influenza virus and in negative-strand RNA viruses generally (4, 5) and has never been observed experimentally. However, we found many sequences in the database that show very strong apparent evidence of homologous recombination. As a rough test for this, we divided the nucleotide sequence of each seg-

* Corresponding author. Mailing address for Michael Krasnitz: Institute for Advanced Study, Einstein Dr., Princeton, NJ 08540. Phone: (609) 734-8048. Fax: (609) 951-4459. E-mail: krasnitz@ias.edu. Mailing address for Raul Rabadan: Institute for Advanced Study, Einstein Dr., Princeton, NJ 08540. Phone: (609) 734-8079. Fax: (609) 951-4459. E-mail: rabadan@ias.edu.

† Supplemental material for this article may be found at http://jvi.asm.org/.

‡ These two authors contributed equally to this work.

▽ Published ahead of print on 25 June 2008.

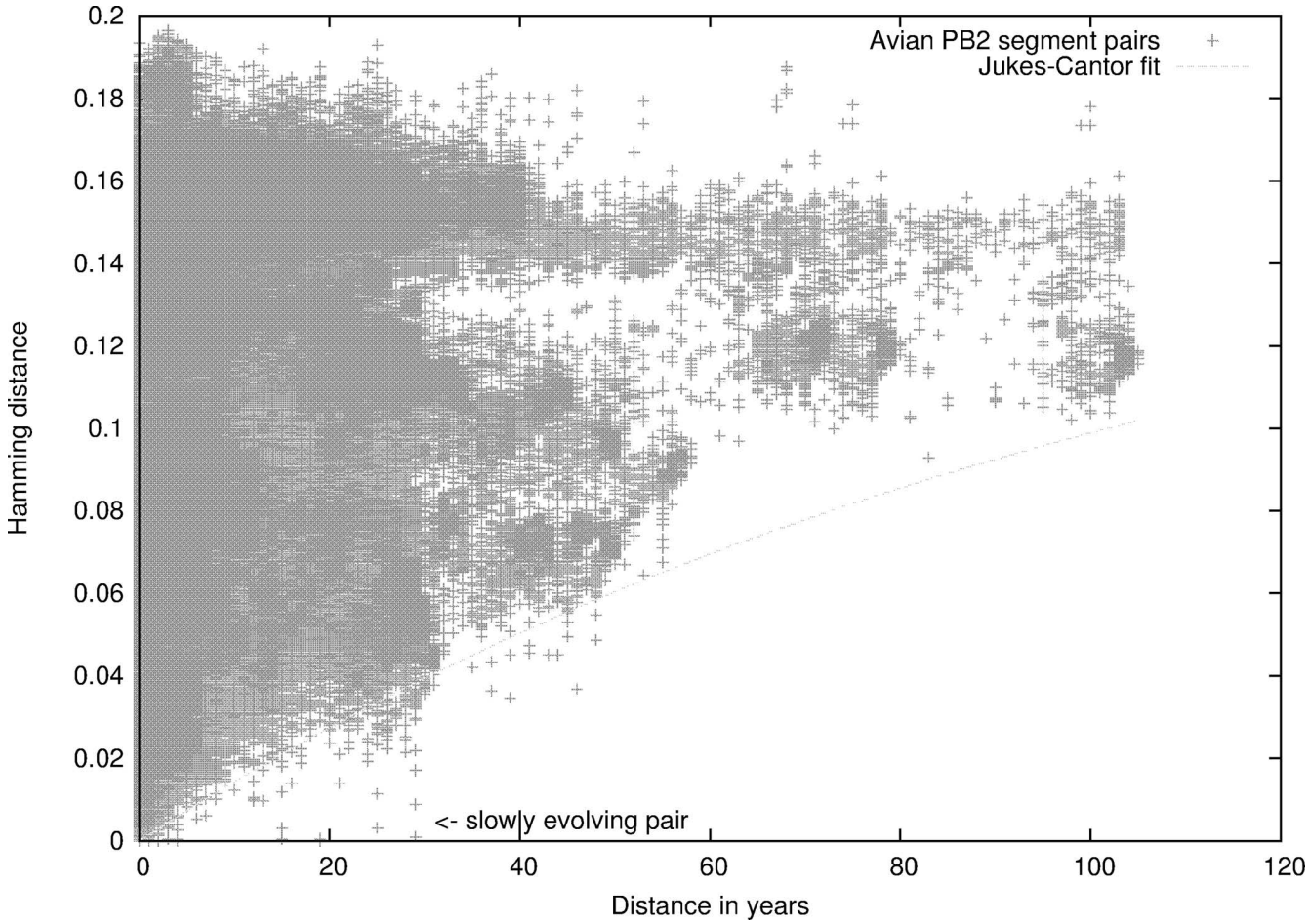

FIG. 1. Hamming distance versus distance in years for PB2 segments from the avian database. Each plus represents a pair of different PB2 sequences of influenza viruses isolated from avian hosts. The *x* axis gives the difference in years between the times of isolation of the two viruses, and the *y* axis gives the Hamming distance between their sequences (number of nucleotide [nt.] differences divided by the length of the segment). The dashed line represents a Jukes-Cantor fit for the expected Hamming distance. An apparently slowly evolving pair is shown.

ment into two equal halves. For each pair of segments, we compared the number of nucleotide differences between them in the first half (i.e., 5′ in the positive strand) with the number of nucleotide differences in the second half. The idea is that if two segments are nearly identical in one part of their sequence but very different in another part, this is strong evidence of homologous recombination, with the divergent parts explained by a recombination event.

In Fig. 2, we plotted a sample comparison for pairs of PB2 segments of viruses isolated from avian hosts. Most points cluster along the diagonal, as would be expected for roughly uniform evolution along the segment with no homologous recombination. However, there are some very significant outliers. For instance, the PB2 sequence of A/shorebird/DE/236/2003(H11N9) differs from that of A/shorebird/DE/231/2003(H9N4) by 6 nucleotides out of the first 1,155 and by 80 out of the second 1,155. Using a null hypergeometric distribution, this gives an extremely low *P* value of 1.6×10^{-18}. As before, these outliers are so extreme that possible corrections allowing for slightly nonuniform evolution along the segment are irrelevant: no model would account for the difference distribution of such a pair as the result of a random fluctuation. Beyond such extreme outliers, a glance at Fig. 2 shows a larger scatter of points lying well off the diagonal "cloud," suggesting that many more sequences are potential apparent recombinants with less significant *P* values.

We performed a systematic search for "recombinant" pairs. For the list of sequences that show strong apparent evidence of homologous recombination events, see Appendix S2 in the supplemental material. With a very conservative *P* value cutoff, we found more than 40 such isolates, again mostly sequences from avian and swine hosts, with many recent H5N1 isolates from Asia. There is a highly significant overlap between the sequences showing evidence of apparent homologous recombination listed in Appendix S2 in the supplemental material and the sequences showing evidence of apparent "frozen evolution" listed in Appendix S1 in the supplemental material. Similar to the "frozen" viruses, isolates often have apparent evidence of homologous recombination in more than one segment, and the overall incidence of recombinant pairs is consistent across the eight segments, allowing for greater statistical power to resolve such pairs in the longer segments.

There have been some historically reported cases of "frozen evolution" in the influenza virus. The most famous of these involves the reemergence in 1977 of H1N1 in the human pop-

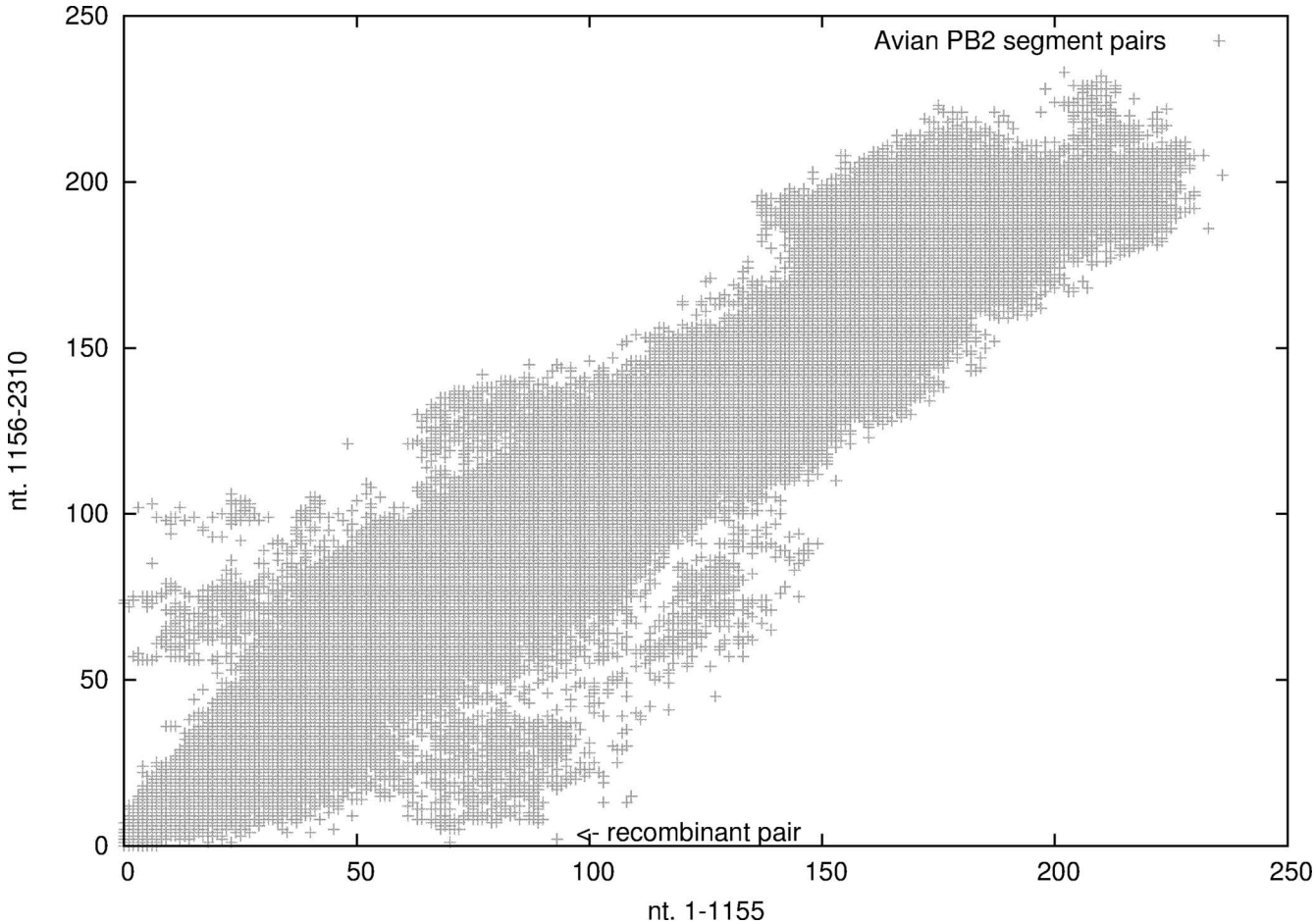

FIG. 2. Number of differences in nucleotides 1156 to 2310 versus the number of differences in nucleotides 1 to 1355 in PB2 segments from the avian database. Each plus represents a pair of different PB2 viruses isolated from avian hosts. An apparently recombinant pair is shown.

ulation after an absence of 20 years. The viruses isolated in the former USSR and China in 1977 were virtually identical in their nucleotide sequences to H1N1 viruses from 1949 to 1950. (We readily detected all previously known "frozen viruses" in our search but omitted them from the list in Appendix S1 in the supplemental material.) In this case, it is believed that the term "frozen" applies literally; these viruses were probably stored in a laboratory for 27 years and then reintroduced into the population in a vaccination experiment gone wrong. There are additional examples involving common laboratory strains PR/34 and WSN, which appeared to reemerge unchanged in humans and camels in Mongolia (1, 23) in the 1980s and pigs in South Korea (7; Seo et al., unpublished) in 2004, respectively. These cases are believed to be explained by escaped vaccines in the former case and stock contamination in the laboratory in the latter.

What can account for the many "frozen" sequences reported here? One possibility is that some interesting biological mechanism is at work. For example, it is possible that the "frozen" viruses are mutating at a slower rate, perhaps because of a more faithful polymerase. To examine this possibility, we searched for amino acid mutations in the polymerase genes (and other genes) common to the "frozen" viruses but were unable to find any such mutations (data not shown). Given the lack of error correction for RNA to RNA polymerase, a mutation that dramatically reduces replication errors does not appear plausible. Another possibility is that these viruses have a much lower rate of replication, perhaps because they persist without replicating within host cells or even in the outside environment, but there is no known latency mechanism for RNA viruses, and the very long times (often decades) elapsed between isolations of nearly identical viruses make this kind of mechanism seem somewhat unlikely (12, 13, 15, 17–21, 24). A recent article argues against the likelihood of influenza virus persistence in the outside environment, such as environmental ice (22). It is important to note that the notion of "evolutionary stasis," which may or may not hold for influenza virus in certain hosts, is not relevant to these results; even viruses that are "static" at the amino acid level are expected to have normal rates of drift in synonymous third-codon nucleotide positions.

We speculate that perhaps the most likely explanation for both of the anomalies reported here is stock contamination in the sequencing laboratories (or wherever the viruses are stored). If the virus stock containing virus A is contaminated with virus B, an experiment supposedly sequencing virus A is actually sequencing virus B, thus resulting in apparent near sequence identity between viruses A and B; if viruses A and B are separated by many years, this will appear as an anoma-

lously low evolutionary rate. If viruses A and B are mixed in the stock, the reverse transcriptase reaction used during sequencing could jump between an A and a B template, resulting in an apparent homologous recombinant. This possibility is consistent with the fact that there is a very significant overlap between the sequences exhibiting apparent slow evolution and those exhibiting apparent homologous recombination; this overlap would be very difficult to explain on biological grounds but is natural if stock contamination has occurred. Also, there is a relative prevalence of old viruses (isolated before 1990) and viruses sequenced by laboratories in Asia, and especially China, among the anomalous sequences; it is tempting to speculate that such differences could reflect differences in laboratory protocols. Along the same lines, nearly all anomalous sequences come from avian and swine hosts; it seems natural to assume that viruses from human hosts are generally handled with greater care (because of the potential public health hazards resulting from their spread) and are thus less susceptible to stock contamination.

If stock contamination is indeed to blame for these anomalies, the results reported here could represent just the tip of the iceberg. This is because we would detect the contamination of viruses A and B only when A and B are sufficiently distant from each other in time of isolation (resulting in a "frozen" virus) and/or nucleotide sequence (possibly resulting in a "recombinant"). It is natural to assume, however, that most contamination events in fact occur between viruses that are relatively close to each other in both time and sequence, resulting in a reported sequence that is wrong but not wrong enough to be detectable by the present methods; this could perhaps account for the "off-diagonal clouds" in Fig. 2. Thus, the present results suggest that an unknown and possibly quite nontrivial percentage of the data in the influenza sequence database might be compromised, and it is our hope that some steps will be taken by the influenza virus research community to address the issue of quality control in the database. One simple, though certainly insufficient, measure would be to regularly resequence the viruses; we expect that in most cases involving apparent "recombinants," a new sequencing assay would result in a different sequence, since it seems unlikely that the reverse transcriptase jumps would occur in the same positions as before. Aside from such detection steps, more should be done in laboratories to prevent contamination from occurring in the first place. If the rapid growth in the influenza virus genome database can be accompanied by addressing these apparent quality control issues, the influenza virus research community will truly be in possession of an invaluable resource.

REFERENCES

1. **Anchlan, D., S. Ludwig, P. Nymadawa, J. Mendsaikhan, and C. Scholtissek.** 1996. Previous H1N1 influenza A viruses circulating in the Mongolian population. Arch. Virol. **141:**1553–1569.
2. **Bikour, M. H., E. H. Frost, S. Deslandes, B. Talbot, and Y. Elazhary.** 1995. Persistence of a 1930 swine influenza A (H1N1) virus in Quebec. J. Gen. Virol. **76:**2539–2547.
3. **Bikour, M. H., E. H. Frost, S. Deslandes, B. Talbot, J. M. Weber, and Y. Elazhary.** 1995. Recent H3N2 swine influenza virus with haemagglutinin and nucleoprotein genes similar to 1975 human strains. J. Gen. Virol. **76:**697–703.
4. **Boni, M. F., Y. Zhou, J. K. Taubenberger, and E. C. Holmes.** 2008. Homologous recombination is very rare or absent in human influenza A virus. J. Virol. **82:**4807–4811.
5. **Chare, E. R., E. A. Gould, and E. C. Holmes.** 2003. Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. J. Gen. Virol. **84:**2691–2703.
6. **Chen, R., and E. C. Holmes.** 2006. Avian influenza virus exhibits rapid evolutionary dynamics. Mol. Biol. Evol. **23:**2336–2341.
7. **Enserink, M.** 2005. Infectious diseases. Experts dismiss pig flu scare as nonsense. Science **307:**1392.
8. **Fitch, W. M.** 1996. The variety of human virus evolution. Mol. Phylogenet. Evol. **5:**247–258.
9. **Fitch, W. M., J. M. Leiter, X. Q. Li, and P. Palese.** 1991. Positive Darwinian evolution in human influenza A viruses. Proc. Natl. Acad. Sci. USA **88:**4270–4274.
10. **Gorman, O. T., W. J. Bean, and R. G. Webster.** 1992. Evolutionary processes in influenza viruses: divergence, rapid evolution and stasis. Curr. Top. Microbiol. Immunol. **176:**75–97.
11. **Lindstrom, S., A. Endo, S. Sugita, M. Pecoraro, Y. Hiromoto, M. Kamada, T. Takahashi, and K. Nerome.** 1998. Phylogenetic analyses of the matrix and non-structural genes of equine influenza viruses. Arch. Virol. **143:**1585–1598.
12. **Marschall, M., A. Helten, A. Hechtfischer, A. Zach, C. Banaschwski, W. Hell, and H. Meier-Ewert.** 1999. The ORF, regulated synthesis, and persistence-specific variation of influenza C viral NS1 protein. Virology **253:**208–218.
13. **Marschall, M., A. Helten, A. Hechtfischer, A. Zach, and H. Meier-Ewert.** 1998. Persistent infection with an influenza C virus variant is dominantly established in the presence of the parental wild-type virus. Virus Res. **54:**51–58.
14. **Nelson, M. I., L. Simonsen, C. Viboud, M. A. Miller, J. Taylor, K. S. George, S. B. Griesemer, E. Ghedi, N. A. Sengamalay, D. J. Spiro, I. Volkov, B. T. Grenfell, D. J. Lipman, J. K. Taubenberger, and E. C. Homes.** 2006. Stochastic processes are key determinants of short-term evolution in influenza A viruses. PLoS Pathog. **2:**e125.
15. **Park, C. H., K. Matsuda, Y. Sunden, A. Ninomiya, A. Takada, H. Ito, T. Kimura, K. Ochiai, H. Kida, and T. Umemura.** 2003. Persistence of viral RNA segments in the central nervous system of mice after recovery from acute influenza A virus infection. Vet. Microbiol. **97:**259–268.
16. **Parvin, J. D., A. Moscona, W. T. Pan, J. M. Leider, and P. Palese.** 1986. Measurement of the mutation rates of animal viruses: influenza A virus and poliovirus type 1. J. Virol. **59:**377–383.
17. **Tobita, K., T. Tanaka, and Y. Hayase.** 1997. Spontaneous excretion of virus from MDCK cells persistently infected with influenza virus A/PR/8/34. J. Gen. Virol. **78:**563–566.
18. **Tobita, K., T. Tanaka, and Y. Hayase.** 1997. Rescue of a viral gene from VERO cells latently infected with influenza virus B/Lee/40. Virology **236:**130–136.
19. **Urabe, M., T. Tanaka, T. Odagari, M. Tashiro, and K. Tobita.** 1993. Persistence of viral genes in a variant of MDBK cell after productive replication of a mutant of influenza virus A/WSN. Arch. Virol. **128:**97–110.
20. **Urabe, M., T. Tanaka, and K. Tobita.** 1993. MDBK cells which survived infection with a mutant of influenza virus A/WSN and subsequently received many passages contained viral M and NS genes in full length in the absence of virus production. Arch. Virol. **130:**457–462.
21. **Wang, M., and R. G. Webster.** 1990. Lack of persistence of influenza virus genetic information in ducks. Arch. Virol. **111:**263–267.
22. **Worobey, M.** 2008. Phylogenetic evidence against evolutionary stasis and natural abiotic reservoirs of influenza A virus. J. Virol. **82:**3769–3774.
23. **Yamnikova, S. S., J. Mandler, Z. H. Bekh-Ochir, P. Dachtzeren, S. Ludwig, D. K. Lvov, and C. Scholtissek.** 1993. A reassortant H1N1 influenza A virus caused fatal epizootics among camels in Mongolia. Virology **197:**558–563.
24. **Zach, A., M. Marschall, and H. Meier-Ewert.** 1999. Influenza C virus persistence depends on exceptional steps in viral RNA synthesis and transport. Arch. Virol. **144:**463–478.