RESEARCH ARTICLE

Quantifying Pathogen Surveillance Using Temporal Genomic Data

Joseph M. Chan,[a] Raul Rabadan[a,b]

Center for Computational Biology and Bioinformatics[a] and Department of Biomedical Informatics,[b] Columbia University College of Physicians and Surgeons, New York, New York, USA

ABSTRACT **With the advent of deep sequencing, genomic surveillance has become a popular method for detection of infectious disease, supplementing information gathered by classic clinical or serological techniques to identify host-determinant markers and trace the origin of transmission. However, two main factors complicate genomic surveillance. First, pathogens exhibiting high genetic diversity demand higher levels of scrutiny to obtain an accurate representation of the entire population. Second, current systems of detection are nonuniform, with significant gaps in certain geographic locations and animal reservoirs. Despite past unforeseen pandemics like the 2009 swine-origin H1N1 influenza virus, there is no standardized way of evaluating surveillance. A more complete surveillance system should capture a greater proportion of pathogen diversity. Here we present a novel quantitative method of assessing the completeness of genomic surveillance that incorporates the time of sequence collection, as well as the pathogen's evolutionary rate. We propose the *q*2 coefficient, which measures the proportion of sequenced isolates whose closest neighbor in the past is within a genetic distance equivalent to 2 years of evolution, roughly the median time of changing strain selection for influenza A vaccines. Easily interpretable and significantly faster than other methods, the *q*2 coefficient requires no full phylogenetic characterization or use of arbitrary clade definitions. Application of the *q*2 coefficient to influenza A virus confirmed poor sampling of swine and avian populations and identified regions with deficient surveillance. We demonstrate that the *q*2 coefficient can not only be applied to other pathogens, including dengue and West Nile viruses, but also used to describe surveillance dynamics, particularly the effects of different public health policies.**

IMPORTANCE **Surveillance programs have become key assets in determining the emergence or prevalence of pathogens circulating in human and animal populations. Genomic surveillance, in particular, provides comprehensive information on the history of isolates and potential molecular markers for infectivity and pathogenicity. Current techniques for evaluating genomic surveillance are inaccurate, ignoring the pathogen's evolutionary rate and biodiversity, as well as the timing of sequence collection. Using sequence data, we propose the *q*2 coefficient as a quantitative measure of surveillance completeness that combines elements of time and evolution without defining arbitrary criteria for clades or species. Through several case studies of influenza A, dengue, and West Nile viruses, we employed the *q*2 coefficient to identify sampling deficiencies in different host species and locations, as well as examine the effects of different public health policies through historical records of the *q*2 coefficient. These results can guide public health agencies to focus resource allocation and virus collection to bolster specific problems in surveillance.**

Received 14 November 2012 **Accepted** 29 November 2012 **Published** 29 January 2013

Citation Chan JM, Rabadan R. 2013. Quantifying pathogen surveillance using temporal genomic data. mBio 4(1):e00524-12. doi:10.1128/mBio.00524-12.

Editor Rino Rappuoli, Novartis Vaccines and Diagnostics

Address correspondence to Joseph M. Chan, jmc2213@columbia.edu.

Despite many therapeutic and epidemiological advances, infectious diseases—the number one cause of death among children—remain directly responsible for roughly 15 million (>25%) of the annual deaths that occur worldwide (1). Of particular concern are emerging infections (EIs) that include novel entities like HIV/AIDS and severe acute respiratory syndrome and previously existing but rapidly spreading diseases like cholera and the plague (2, 3). Zoonosis is a rich source of such EI transmission into human hosts (4), suggesting that pathogen surveillance of animals, as well as humans, is an important method of early detection of potential outbreaks and of capturing the entire biodiversity of a pathogen population at any given place and time.

Clinical, serologic, and genomic surveillance systems serve as invaluable tools for detecting early outbreaks, determining the genetic variation of a population, improving vaccine design, and evading antibiotic resistance. In the case of influenza A virus, coordinated global efforts like those of the World Health Organization (WHO) identify cases of influenza-like symptoms, conduct serology studies, and sequence viral isolates (5). Despite such efforts, there remain areas of sparse data, particularly in potential tropical influenza hot spots like India, Africa, or South America. Such sampling bias in strain selection can skew the predicted dominant virus used in annual vaccine design (6, 7).

Similarly, influenza virus surveillance has historically centered on human influenza cases but ignored animal hosts, despite the importance of zoonosis in influenza virus transmission. One particular animal host that deserves special attention is poultry, which has played a crucial role in the transmission of highly pathogenic avian influenza (HPAI) virus H5N1 to humans. First reported in Hong Kong in 1997 and 2003, this virus has spread

quickly from waterfowl to chickens, crows, pigeons, and other birds in Europe, Africa, and particularly Asia, leading to the deaths by infection and forced culling of millions of birds and resulting in appreciable economic losses (8). Concurrently, sporadic infections among humans and other mammals have claimed, as of January 2012, a high rate of 340 deaths out of 578 confirmed human infections since 2003 (9).

At present, no human-to-human transmission of HPAI virus has been documented. However, scientists have recently discovered the set of mutations that enable the transmission of avian influenza virus between ferrets (10), the animal model most closely mimicking human pathogenesis because of shared host cellular receptors for viral attachment (11). These developments, combined with the virus's atypically high rate of mortality and apparent resistance to oseltamavir (12), have raised worldwide concerns about whether our current avian influenza virus surveillance is sufficient to identify wild strains that have the potential for mammalian adaptation.

The swine is another animal that demands adequate surveillance (13). Researchers have observed that epithelial cells in the pig trachea contain sialyloligosaccharides SAα2,3 and SAα2,6, which are unique host determinants for birds and humans, respectively (14). This feature confers tropism in pigs on both avian and human influenza viruses (15), allowing swine to serve as mixing vessels for antigenic shift, in which different strains infecting the same host can reassort RNA segments to create a novel viral strain (16). Such reassortments have engendered at least two major human pandemics in the 20th century, in 1957 and 1968 (17). The most recent pandemic of swine origin influenza virus (SOIV) in 2009, in particular, was a reassortant between Eurasian and North American swine lineages, and although the virus most likely came from a pig, a definitive geographic origin has yet to be determined (13, 18). Interestingly, the closest ancestors of the 2009 pandemic virus were related to viruses isolated from swine more than a decade prior (13, 19), suggesting that relevant strains circulating in swine herds have escaped detection because of inadequate sampling (18).

In response to the dearth of avian and swine influenza virus isolates, invaluable programs like the Global Initiative on Sharing Avian Influenza Data (GISAID) and the PREDICT program sponsored by the U.S. Agency for International Development Emerging Pandemic Threats were established to focus the monitoring of circulating strains on wildlife, as well as to collect and make public clinical, epidemiological, and molecular data on influenza virus (20, 21). Despite such initiatives, many countries with the greatest number of poultry or pig stocks do not proportionately contribute avian and swine influenza virus sequences to public repositories (7).

Currently, there is no standard quantitative measurement of the appropriate level of genomic surveillance of a given pathogen. One immediate approach might be to consider the absolute number of sequenced isolates, but this method fails to account for the diversity and time information of the population sampled. A comparatively large number of sequences may be insufficient to capture high genetic diversity in a pathogen population. Another possible strategy appeals to clustering techniques. After representing sequences as points in a mutational landscape, a highly clustered structure could potentially reflect highly biased sampling. Once again, this method ignores time and can be confounded by evolutionary processes like bottlenecks. A final tactic might be to measure the genetic diversity within a sample of the pathogen. Phylogenetic reconstruction and tallying of species richness is one method that characterized the subclades of avian H5N1 (8), but such analyses are cumbersome, with arbitrary boundaries for classifying clades. Techniques more grounded in information theory include Shannon's entropy and Rao's quadratic entropy (22), but high diversity in a sample does not necessarily correlate with high surveillance either.

In this paper, we propose a readily interpretable and computable quantitative measurement for genomic surveillance of a pathogen that directly accounts for the number of isolates, the evolutionary rate, and the time of sample collection without the need to define arbitrary clades or species or the need for a full phylogenetic reconstruction. This measure ranks a surveillance system as more complete if it is able to capture a greater proportion of the pathogen's diversity. We apply this measure to influenza virus and compare the surveillance of different influenza virus strains in different hosts and geographic regions. We find that, compared to human seasonal strains, sampling is indeed substantially lower for swine H1N1 and avian non-H5N1 influenza virus, historically overlooked strains despite their pandemic potential. We also find that avian H5N1 influenza virus surveillance in the WHO transmission zones of northern and western Africa; eastern, southern, and southeastern Asia; and eastern, southwestern, and northern Europe is high and may potentially serve as an effective early warning system given a list of genetic determinants of mammalian adaptation. Avian H5N1 influenza virus surveillance in North America, however, is much less comprehensive. We similarly apply our methodology to other RNA viruses and show inadequate surveillance of both dengue and West Nile viruses. Finally, we perform a comparative analysis of the $q2$ coefficient and other methods, particularly phylogenetic and clustering alternatives, and find that the $q2$ coefficient produces similar results with negligible computation time.

RESULTS

Measurement of surveillance: the *q*2 coefficient. We propose a quantitative measure of pathogen surveillance that reflects both the evolutionary rate and the biodiversity of a given population. For each host and subtype, sequences were first temporally ordered from the oldest sequence to the newest. For each sequence, we identified the sequence from the past with the highest homology and recorded its genetic distance, such that for N sequences, we compiled a vector of N distances. We defined the surveillance measurement $q2$ coefficient, a measurement between 0 and 1, as the proportion of sequences within genetic distance R of the most closely related ancestor.

$$q2 = \frac{Genetic\ Distance < R}{N}$$

Given a viral evolutionary rate of μ substitutions per site per year, $R = t \times \mu$ is thus the expected proportion of mutations that have accumulated over the span of t years. In this report, we examine short time periods on the order of a decade and do not expect the true genetic distance to diverge greatly from the Hamming distance. However, we calculate the $q2$ coefficient by using a number of distance methods, including Hamming distance, Jukes-Cantor, Kimura, Tamura, Tajima-Nei, Hasegawa, and Nei-Tamura, and find little significant change (see Results). If the surveillance of a pathogen in a given time and place is sufficient, we

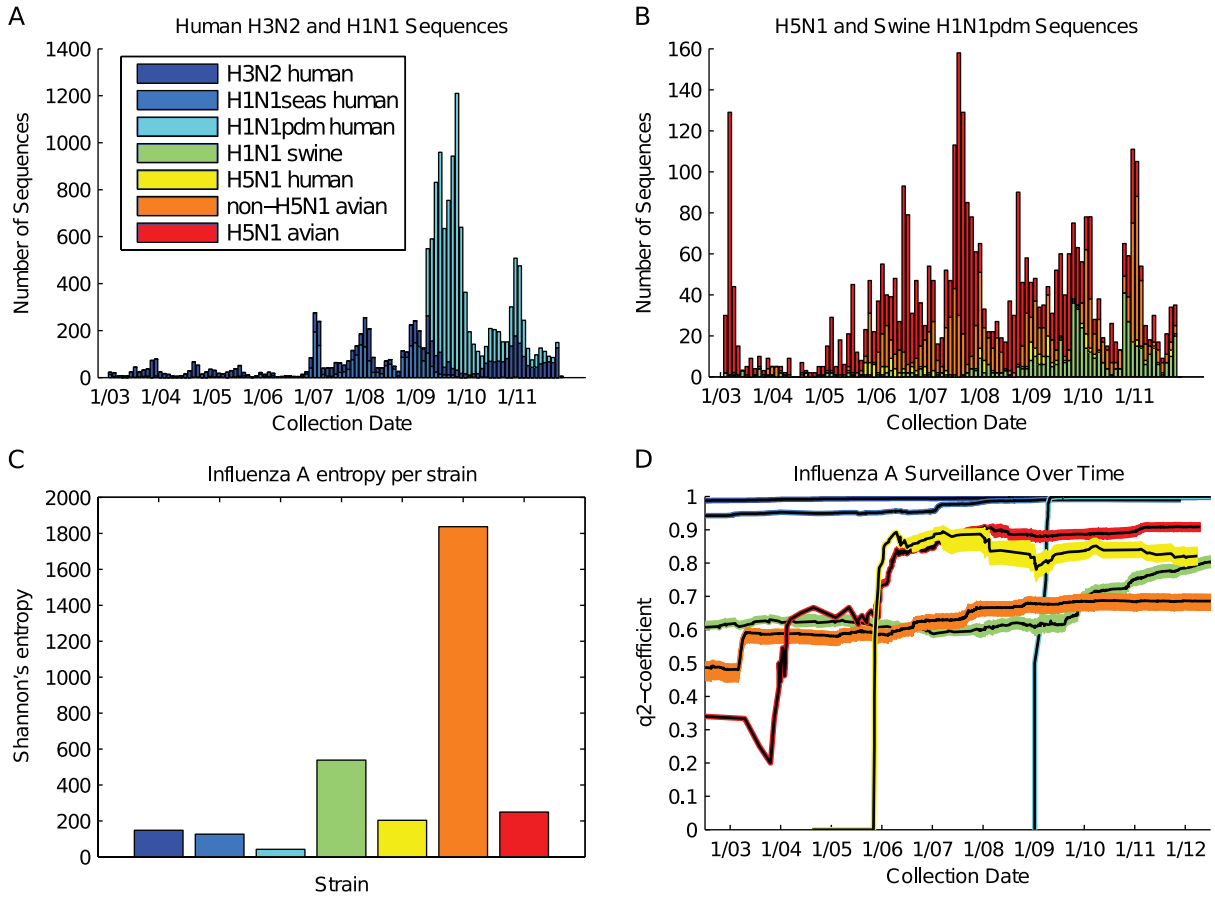

FIG 1 Surveillance of influenza A virus. (A) Numbers of human H3N2 and H1N1 sequences over time. (B) Numbers of H5N1 and swine H1N1pdm sequences over time. (C) Measurement of influenza A virus genetic diversity by entropy. (D) Measurement of influenza A virus surveillance by the $q2$ coefficient over time. The width of each plotted line denotes the interval of $q2$ based on the 95% HPD of the evolutionary rate. $n = 7{,}083$ H3N2 human (dark blue), 2,567 H1N1 human (blue), 11,626 H1N1pdm human (cyan), 878 H1N1 swine (green), 1,193 H5N1 avian (yellow), 158 H5N1 human (orange), and 3,418 non-H5N1 avian (red) influenza virus sequences.

would therefore expect a maximal number of viral isolates to have a closest ancestor in the sequence database with a sequence identity of greater than $1 - R$. For the q coefficient, we chose to use $t = 2$ years since antigenic variant strains of influenza virus emerge and become predominant over a period of roughly 2 to 5 years (23); however, any value of t can be used. One consideration is that substitution rates can vary over time and across different lineages. To incorporate such variability into our calculation of the $q2$ coefficient, we used the 95% highest posterior density (HPD) interval of the evolutionary rate collected from the literature (see Table S1 in the supplemental material) (24–26).

A major motivation behind use of the $q2$ coefficient is the ability to evaluate the surveillance of different strains, hosts, and geographic regions. Sampling during variable windows of time, however, can compromise an appropriate comparison. For example, a high $q2$ coefficient derived from a local weeklong outbreak should not be extrapolated to surveillance during an entire season. For any comparisons of surveillance over a span of years, we therefore exclude from analysis any groups not sampled for more than a month.

Influenza virus surveillance. Complete coding sequences of hemagglutinin (HA), the major antigenic segment of influenza virus, were collected from the NCBI (National Center for Biotechnology Information) and GISAID databases and multiply aligned. To compare the levels of surveillance of different strains, we separately considered sequences for human (H3N2, seasonal H1N1 pre-2009, SOIV H1N1 post-2009, and H5N1), avian (H5N1 and non-H5N1), and swine H1N1 influenza viruses. Figure 1A and B depict the absolute number of sequences isolated across time as a first measure of surveillance. However, we wanted a measure of surveillance that would take into account the evolutionary rate of the virus, as well as the biodiversity of the viral pool. Calculation of Shannon's entropy in Fig. 1C is an effective measure of genetic diversity but, by itself, is no measure of surveillance.

Toward a measure that synthesizes both the absolute number of isolates and the evolutionary rate, we calculated the $q2$ coefficient as a function of time for each influenza virus strain in Fig. 1D as a representation of surveillance history. We also tabulated their final $q2$ coefficients (see Tables S2 and S3 in the supplemental material) and mapped these values among WHO transmission zones (27) in Fig. 2 (and Fig. S2 in the supplemental material) to denote the present state of surveillance around the world. We showed that our $q2$ coefficient calculations did not drastically change for different genetic distances (<0.035 difference in the $q2$ coefficient).

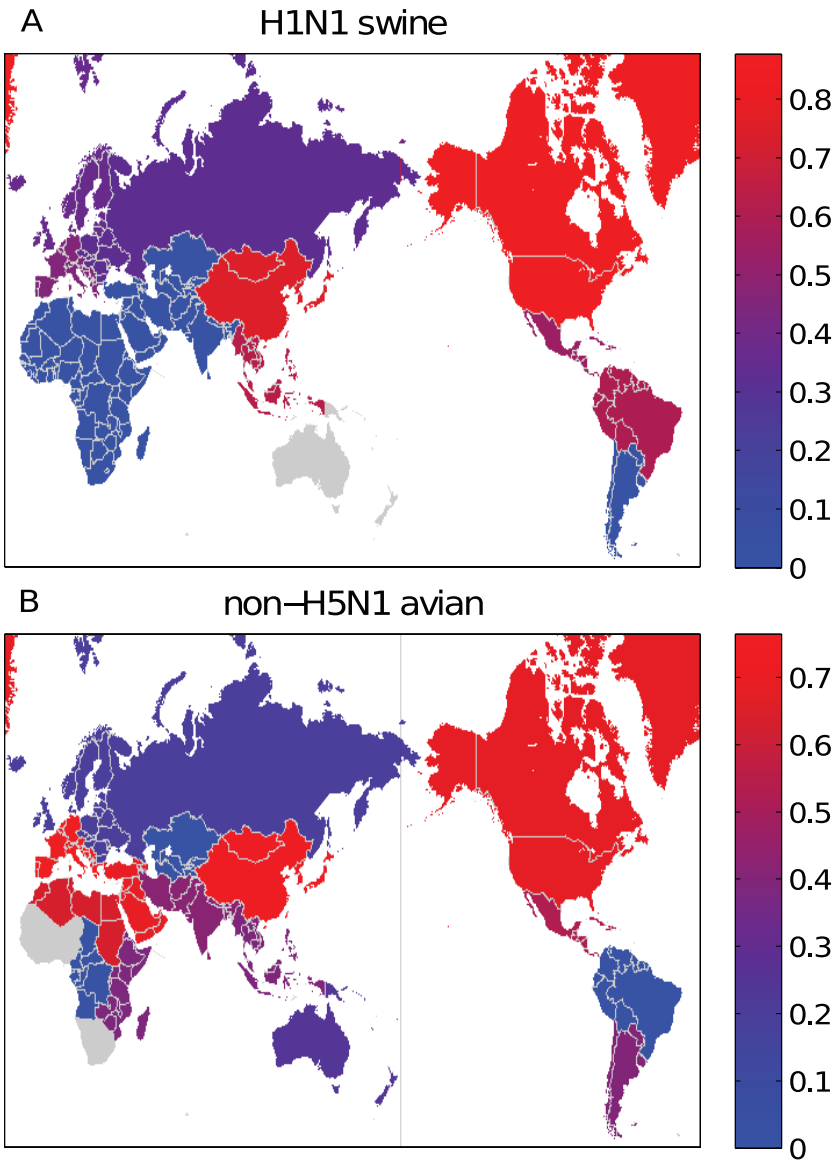

FIG 2 Global map of influenza A virus surveillance in animal hosts by transmission zones. The strains depicted include swine H1N1 (A) and non-H5N1 (B) avian influenza virus strains. Each zone is colored in the blue-to-red spectrum to indicate the *q*2 coefficient. Gray areas denote regions with viruses isolated over a span of no more than 1 month.

Analysis of each strain reflects a different state of geographic and host surveillance. Overall, surveillance of human seasonal strains was high; the *q*2 coefficient values of both human seasonal H3N2 and seasonal H1N1 viruses, which had been sampled and sequenced long before 2003, approached 1 from 2003 to the present (Fig. 1D). Clustering of sequences by transmission zone began to uncover weakness in the surveillance of seasonal H3N2, particularly in central Asia (see Fig. S1A and S2A in the supplemental material). In addition to central Asia, several parts of Europe, western Asia, and especially central Africa, yielded lower *q*2 coefficients of seasonal H1N1 (see Fig. S1B and S2B in the supplemental material). As a testament to the global response to the pandemic, human SOIV H1N1 shot up to a *q*2 coefficient of 1 shortly after its arrival in March 2009 in all transmission zones except central and southern Africa and central Asia (Fig. 1D; see Fig. S1C and S2C in the supplemental material).

In contrast, despite a moderate increase in the number of sequences following H1N1pdm's arrival (Fig. 1B), the *q*2 coefficient of classical H1N1 swine isolates has lagged much further behind (Fig. 1D) and is consistently high only in the eastern Asian (*q*2 coefficient = 0.753, Hamming distance) and North American (*q*2 coefficient = 0.877, Hamming distance) transmission zones (Fig. 2A; see Fig. S1D in the supplemental material). These findings suggest that surveillance of pigs is still not enough to capture the high biodiversity of swine flu, as indicated by entropy (Fig. 1C).

Analysis of H5N1 influenza virus describes the effects of implementing international sequencing initiatives like GISAID. Immediately following the second outbreak of HPAI virus among humans in 2003, the *q*2 coefficient of H5N1 in both human and avian hosts rose to approximately 0.7 to 0.85. With the establishment of GISAID in 2006, both human and avian H5N1 influenza virus *q*2 coefficients increased further to high levels of around 0.9, affirming the effectiveness of the global consortium. Beyond 2007, however, the surveillance of human H5N1 influenza virus has waned compared to that of avian H5N1 influenza virus (Fig. 1D). This discrepancy is most likely due to the fact that human HPAI cases represent a subsampling of avian H5N1 influenza virus genotypes, as reflected by the slight drop in human H5N1 entropy compared to that in birds (Fig. 1C).

Within H5N1, there is biased sampling in different transmission zones. Following particularly large outbreaks, H5N1 human surveillance is high, with *q*2 coefficients of roughly 0.9 in northern Africa and eastern and southeastern Asia. Over time, the *q*2 coefficient has decreased in eastern and southeastern Asia most likely because of a decline in the number of sporadic introductions into the local human populations compared to that in northern Africa, specifically Egypt (see Fig. S1E and S2D in the supplemental material). H5N1 avian influenza virus surveillance is high in northern and western Africa; eastern, southeastern, and southern Asia; and eastern, southwestern, and northern Europe. On the other hand, the *q*2 coefficient indicates less avian H5N1 influenza virus surveillance in North America. In the United States, for example, only until 2006 were the reporting and tracking of H5 in birds mandated by the USDA (28). The smaller push for reporting stems from the low pathogenicity displayed by North American avian H5N1 influenza virus strains, which have antigenic and genetic differences from the Asian HPAI virus lineage (see Fig. S1F and S2E in the supplemental material) (29).

These results indicate that H5N1 influenza virus surveillance of

avian hosts is much more complete than H1N1 surveillance of swine. However, potential zoonotic transmission from other avian strains is possible, as well. Performance of the same analysis of non-H5N1 avian influenza virus strains yielded a low *q*2 coefficient beginning at 0.5 in 2003 and plateauing at a level just below 0.7 in spite of GISAID (Fig. 1D) and the resulting increase in sequenced isolates (Fig. 1B). This finding potentially reflects the extremely high genetic diversity of influenza A virus in its natural reservoir (Fig. 1C) that is not fully captured by current surveillance systems. Transmission zone analysis of non-H5N1 avian influenza virus strains indicates that non-H5N1 surveillance is concentrated in southwestern Europe, the Central American-Caribbean region, North America, northern Africa, and eastern and southeastern Asia (Fig. 2B; see Fig. S1G in the supplemental material).

Calculation of the *q*2 coefficient for other complete coding segments of influenza virus was also performed. Since observed differences in sequences should reflect the evolutionary rates of each segment, any differences in the *q*2 coefficient should reflect differences in sampling alone. For H3N2, H1N1pdm, and seasonal H1N1, the *q*2 coefficient exhibited little change among different segments (differences of <0.02) and moderate change for other strains. The largest change in the *q*2 coefficient was 0.186 between swine H1N1 HA and PB2. Despite such differences between segments, results showed that generally surveillance of human H3N2, H1N1pdm, and seasonal H1N1 across all segments surpassed that of human H5N1, avian H5N1, swine H1N1, and non-H5N1 avian influenza viruses (see Tables S2 and S3 in the supplemental material).

Dengue and West Nile virus surveillance. As a proof of concept, we have also shown that the *q*2 coefficient can be used to monitor surveillance efforts for other RNA viruses, such as dengue virus and West Nile virus. Here, we focus on the *env* gene of these viruses, as it encodes the longest of the structural proteins, which is prevalently sequenced for subtyping and has the best-documented evolutionary rates of all flavivirus proteins (25, 26, 30, 31).

Calculation of the *q*2 coefficients of dengue virus subtypes 1, 2, and 3 depicts poor surveillance in general (Fig. 3; see Fig. S3 and Table S4 in the supplemental material). These sampling gaps may reflect the current limited funding and staff in many tropical countries around the world (32). The incorporation of other genetic distances besides Hamming distance produced a <0.118 change in the *q*2 coefficient of dengue virus.

The *q*2 coefficient of West Nile virus in the United States, on the other hand, was higher in the first few years after its introduction into the United States (see Fig. S4 and Table S5 in the supplemental material). However, despite the implementation of early warning systems for West Nile virus, by sampling dead birds (33) and mosquito populations (34), the *q*2 coefficient beyond 2003 dropped over time, even with the addition of multiple isolates within a short period of time. This decline in the *q*2 coefficient coincided with a rapid population growth of West Nile virus in the United States during 2002 and 2003 spurred by the establishment of the WN02 strain marked by a V159A mutation in its E protein (35). This analysis suggests that current surveillance is being outpaced by the growing diversity of the virus in the New World. Other genetic distances besides the Hamming distance produced a <0.094 change in the *q*2 coefficient of West Nile virus.

Comparison to phylogenetic methods. One may wonder if other alternative methods that account for pathogen evolution may suffice to characterize the genetic surveillance of a pathogen. Phylogenetics has been used in many studies to characterize pathogen surveillance qualitatively without producing a quantitative measure of sampling completeness (36). A possible phylogenetic analogue to the *q*2 coefficient might entail the reconstruction of a tree based on available sequences and measurement of the distribution of branch lengths. The true distance between two isolates, A and B, is represented by the sum of their patristic distances, d_A and d_B, which are the branch lengths from each respective sequence to their common ancestral node. Sequences are time ordered, however, and if we assume an approximate molecular clock, then $d_A < d_B$ given that sequence A occurs before sequence B. An estimate of distance is then the larger patristic distance. A parallel to our *q*2 coefficient would predict high surveillance to correspond to a maximal number (#) of patristic distances *d* to their closest ancestor in the past less than 2 years as follows:

$$p2 = \frac{\#\ branch\ lengths\ d < 2\ years}{Total\ \#\ branch\ lengths\ d}$$

Moreover, homogeneity of surveillance can be confirmed if branch lengths *d* have low variance.

Phylogenies can be divided into those that are distance based and those that are character based. Since the *q*2 coefficient readily incorporates different genetic distance methods, it is equivalent to any *p*2 coefficient calculated from distance-based trees. On the other hand, character-based trees, including maximum-likelihood and Bayesian inference methods, incorporate site heterogeneity by considering one character (a site in the alignment) at a time to reconstruct a tree (37); moreover, Markov chain Monte Carlo (MCMC) methods like BEAST (38) can incorporate relaxed clock rates. The *q*2 coefficient does not take into account either site or clock rate heterogeneity.

To determine the impact of site and clock rate heterogeneity in quantifying surveillance completeness, we calculated the *p*2 coefficient of the human H5N1 HA data set of 158 sequences by using BEAST (see Materials and Methods). We accounted for site heterogeneity by using the gamma model (39) and reconstructed trees under both strict and relaxed molecular clocks. We calculated the *p*2 coefficients to be 0.848 (0.740 to 0.917) and 0.860 (0.721 to 0.911) for the strict and relaxed clocks, respectively. Given our *q*2 coefficient of 0.821 (0.795 to 0.848) for human H5N1 HA, we concluded that incorporating site heterogeneity and a relaxed molecular clock did not make a significant difference.

While these phylogenetic techniques can examine the fit of a number of evolutionary models, they suffer from problems of robustness. For example, tree topology can be highly unstable; the addition or deletion of a single sequence can radically restructure the tree. Moreover, different methods of phylogenetic inference, such as maximum likelihood or Bayesian inference, can lead to variable results, rendering interpretation of surveillance complicated. Finally, computation time, particularly for BEAST, can be very expensive; for data sets of more than 1,000 sequences, several weeks may be needed for the MCMC to converge to a stable tree solution. In our analysis of 158 human H5N1 HA sequences, *p*2 coefficients needed days of computation to complete, whereas *q*2 coefficient analysis was finished in a matter of seconds.

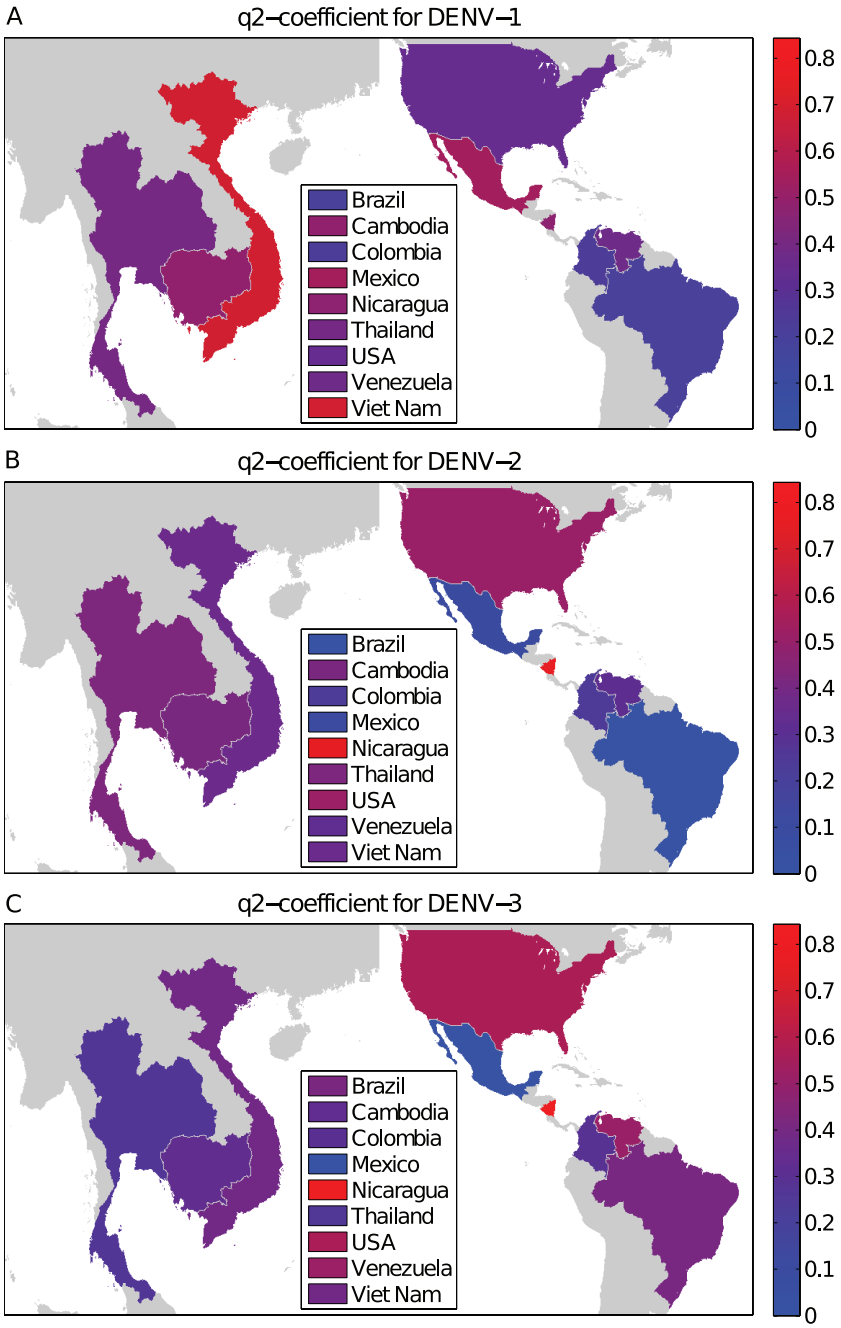

FIG 3 Global map of dengue virus (DENV) surveillance in different countries. The strains depicted include DENV-1 (A), DENV-2 (B), and DENV-3 (C). Each zone is colored in the blue-to-red spectrum to indicate the $q2$ coefficient. Gray areas denote regions with viruses isolated over a span of no more than 1 month.

Comparison to clustering methods. Another possible surveillance measurement characterizes the cluster structure of isolates. In an ideal situation, a well-sampled population of sequences separated by genetic distance would be represented by points densely and homogeneously spread across a continuum. Therefore, clustering techniques such as hierarchical, k-means, or expectation-maximization clustering can be used to ascertain how poorly sampled a pathogen is on the basis of the number of clusters in a data set. Bar coding is an alternative strategy based on the field of persistent homology that identifies topologically invariant clusters in cloud data; in particular, it calculates the b_0 Betti number, the number of connected components in a set of simplicial complexes constructed from sequences at different filtration Hamming distances (see Materials and Methods) (40). A lower b_0 would indicate better sampling.

As a comparison to the $q2$ coefficient, we applied bar coding to determine the b_0 values of different influenza virus strains at filtration Hamming distances $\boldsymbol{\varepsilon}$ ranging from 0 to 200 (Fig. 4A). This threshold $\boldsymbol{\varepsilon}$ is analogous to the R threshold of the $q2$ coefficient: 2 years of influenza virus evolution equivalent to $2 \times (5 \times 10^{-3}) = 1\%$ genetic distance. We therefore considered b_0 at a Hamming distance of 1% of the length of HA, or roughly 17 nucleotides, to be appropriate for comparison to the $q2$ coefficient (Fig. 4B). For the most part, the $q2$ coefficient and b_0 are inversely correlated. For example, the low b_0 and high $q2$ coefficient of human H3N2, seasonal H1N1, and H1N1pdm indicate good surveillance. Non-H5N1 avian influenza virus has an extremely high b_0 and a low $q2$ coefficient. It should be noted that the non-H5N1 avian influenza virus data set contains 15 different HA subtypes, as opposed to all of the other single-subtype HA data sets considered. Thus, b_0 may need to be normalized by the expected number of HA subtypes. Nevertheless, b_0 for non-H5N1 avian influenza virus remains high even with normalization. However, as with $p2$, calculation of b_0 demands substantial computing power and time on the order of hours.

Comparison to diversity metrics. Other alternative methods of gauging sampling bias may exist. For instance, there are many diversity metrics in ecology, including Chao's estimate of asymptotic species richness (41), and in information theory, techniques like Shannon's entropy, which was applied in Fig. 1C. In particular, it is well known that the empirical Shannon entropy of a sample underestimates the entropy of the true distribution. Bootstrapping techniques can be used to estimate the bias produced by a low number of sequences. This method is attractive for its ability to measure surveillance by assaying the diversity of a pathogen directly. However, like the phylogenetic and bar coding methods considered, implementation for a large number of trials is substantially more difficult, computationally expensive, and time-consuming.

DISCUSSION

In this paper, we offer a quantitative method of measuring the quality of surveillance of a particular pathogen that reflects its evolutionary rate and genetic diversity. The $q2$ coefficient repre-

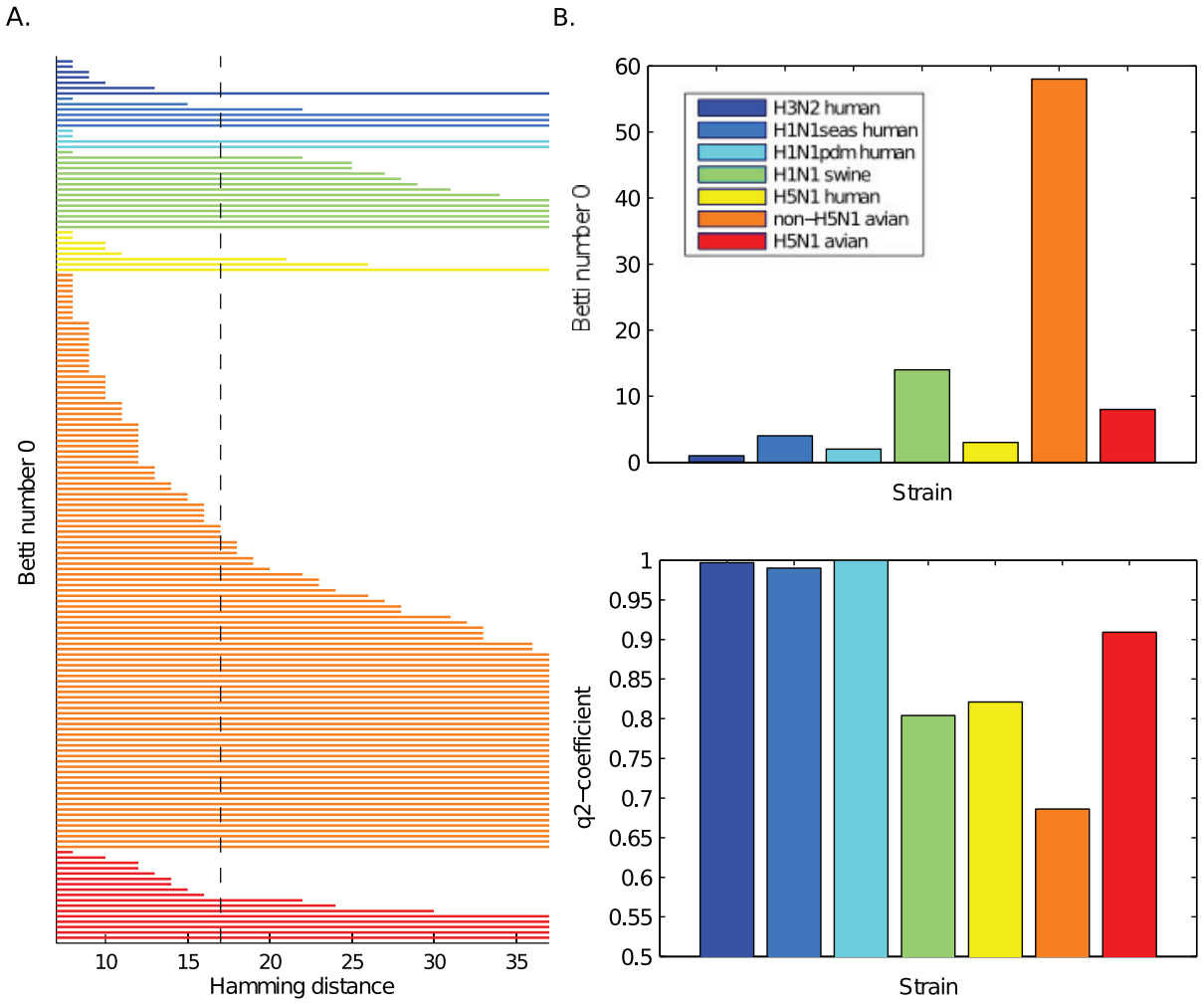

FIG 4 Persistence homology analysis of influenza virus. (A) Bar coding plots of b_0 for each strain of influenza virus. b_0, or the number of simplicial complexes, can be determined by counting the bars present at each filtration Hamming distance. A lower b_0 value indicates better surveillance, although this number may be higher in a setting of bottleneck effects. (B) Comparison of persistence homology and the $q2$ coefficient applied to influenza virus. The top bar plot displays b_0 at a Hamming distance of 17, roughly 1% of the length of HA. The bottom bar plot displays $q2$ coefficients, with a threshold of 2 years of influenza virus evolution, or $2 \times (5 \times 10^{-3}) = 1\%$.

sents an easily interpretable quantitative measure of genomic surveillance that does not rely on a full phylogenetic reconstruction. A number between 0 and 1, the $q2$ coefficient reflects the fraction of isolates with a closest isolate in the past within 2 years of evolution. However, in its simplicity, the $q2$ coefficient does not capture the complexity of evolutionary processes such as variable evolutionary rates, recombination, reassortment, and population genetic effects. We decided on the $q2$ coefficient instead of other equally valid metrics ($Q1$, $q5$, or $q100$) because 2 years represents the median influenza A virus vaccine update time. However, different metrics may be applicable, depending on the specific aim of the surveillance program.

We applied this measure to multiple influenza virus strains and found that current surveillance is generally complete for humans, particularly H1N1pdm, H3N2, and seasonal H1N1, but poor for other hosts, particularly swine. Indeed, the calculated $q2$ coefficient of swine H1N1 is most likely an overestimate, considering that influenza virus evolves more slowly in swine than in humans because of a lesser degree of immune selection (42). This finding reaffirms the need for increased surveillance of swine, which can serve as mixing vessels for pandemic strain selection.

Compared to human H3N2 and H1N1 surveillance, human and avian H5N1 influenza virus surveillance is lower at $q2$ coefficients of 0.838 and 0.923, respectively. Although these the $q2$ coefficients are already high, the difference in the $q2$ coefficient may suggest that there are strains of H5N1 circulating in bird and even possibly human hosts that remain undetected. Given the recent discovery of molecular determinants that enable ferret-to-ferret transmission of H5N1 virus (10), it is all the more important to increase viral isolation to match or exceed the levels of human H3N2 and H1N1 surveillance.

In our analysis, we also noted geographic disparities; a preponderance of H5N1 avian influenza virus isolates derived from northern and western Africa; eastern, southeastern, and southern

Asia; and eastern, southwestern, and northern Europe. Although North America's surveillance of non-H5N1 avian influenza virus sequences is trending upward, it is particularly deficient in the sampling of H5N1 avian influenza virus. Even though only low-pathogenicity avian influenza virus has been discovered in North America, it is clear that current surveillance is not sufficient to capture the complete diversity of even this H5N1 population.

One may wonder whether factors beyond the completeness of surveillance, such as natural selection, may influence the $q2$ coefficient. For instance, bottlenecks and selective sweeps can reduce the diversity of a pathogen population, thus increasing the $q2$ coefficient. However, it is important to note that in such cases, the $q2$ coefficient behaves appropriately, for it simply reflects surveillance completeness. If only a few sequences are enough to capture the reduced biodiversity following a bottleneck, the $q2$ coefficient will be close to 1.

Another possible confounding effect stems from selection bias in the submission of sequences to public databases. We account for any possible duplication of sequences by considering only unique sequences from each date and location. Beyond this safeguard, we acknowledge that the $q2$ coefficient may fall prey to selection bias. However, without prior knowledge of such biases, it would be difficult for any method to address these confounders satisfactorily.

We compared the $q2$ coefficient to other possible quantitative methods, including those based on phylogenetics and clustering. A drawback common to many of these surveillance methods is that they are slow, computationally expensive techniques that are not explicitly designed to capture pathogen diversity. For this reason, we developed the $q2$ coefficient, which is readily computable and captures surveillance at any given time point. Applying the $q2$ coefficient revealed deficient sampling of swine H1N1 and nonavian H5N1, dengue, and West Nile viruses. These results bear great potential to guide the future allocation of energy and resources for gathering viral isolates. We also foresee further ecological applications of the $q2$ coefficient as an effective measure of asymptotic species richness without relying on arbitrary criteria for defining species, unlike classic techniques like Chao's estimate. This feature can be particularly valuable for assessing the number of additional samples needed to represent all of the species of an organism in metagenomic studies. In conclusion, the scientific community as a whole must improve its surveillance networks and share information for the advancement of scientific inquiry and public health interventions, and we offer the $q2$ coefficient as a means of evaluating the effectiveness of such efforts.

MATERIALS AND METHODS

Sequence collection, annotation, and alignment. Complete coding sequences of all influenza virus coding segments were collected from the NCBI influenza virus resource (43) and the GISAID database (20) downloaded as of October 2012 for human, avian, and swine hosts. Complete coding sequences for the envelope (*env*) gene of dengue virus subtypes 1, 2, and 3, as well as West Nile virus, were collected from the Virus Pathogen Resource and Broad Institute downloaded as of May 2012. All sequences were filtered for year and month information. If no day information was provided, isolation was assumed to have occurred on the 15th of the month. To avoid the effect of depositing duplicated sequences, only unique sequences were considered for each date and location. Sequences were then aligned using ClustalW2 v. 2.0.12, using default parameters, and then manually curated. Influenza virus sequences were annotated by transmission zones, defined by the WHO as geographically contiguous regions with similar peaks in the influenza season (27). All alignments and the code used are available upon request.

Shannon's entropy as a measurement of genetic diversity. Our measurement of pathogen surveillance incorporates the evolutionary rate, which contributes to population diversity. To measure genetic diversity directly, we chose to employ Shannon's entropy, a popular and intuitive measure that avoids cumbersome phylogenetic reconstruction, and applied it to the distribution of alleles at each nucleotide position within an alignment. This calculation recovers the number of bits of information per base. Assuming the independence of each position, the total entropy of a population is the sum of the entropies of each base position of the alignment. To correct for bias stemming from a variable number of isolates (n), we estimate Shannon's entropy (H) for a given base position on the basis of the following algorithm of Miller et al., where m is the number of alleles (44):

$$H = \frac{m-1}{2n} - \sum_{i=1}^{n} p_i log_2 pi$$

Phylogeny as a measurement of surveillance. Sequences were analyzed by using BEAST v1.7.4, a Bayesian MCMC approach, to sample time-structured evolutionary trees from their joint posterior probability distribution. Because of computational and time limits, we restricted our phylogenetic analysis to human H5N1, as BEAST analysis of data sets of over a thousand sequences can take weeks to converge to a stationary condition. This data set was analyzed by using an exponential-growth coalescent as a tree prior with an HKY + Γ model of nucleotide substitution. Relaxed and strict molecular clocks were employed. For the strict (relaxed) clock, 20 independent runs of 750,000 (1,500,000) steps each were performed, compared for convergence, and combined, with a burn-in of 150,000 (500,000) steps from each. Each run finished after an average of 32.47 (19.14) h using one 3.00-GHz Intel Xeon CPU with 12 GB of random-access memory. A single maximum clade credibility (MCC) tree was created from each set of simulations with the average length determined for each branch. We found the corresponding evolutionary rates to be 5.11×10^{-3} (range, 4.58×10^{-3} to 5.62×10^{-3}) and 5.39×10^{-3} (range 4.39×10^{-3} to 5.80×10^{-3}), similar to the evolutionary rate we used for influenza virus HA analysis. The 95% HPD of each branch length of the MCC tree was used to set a confidence interval for each $p2$ coefficient.

Persistence homology as a measurement of surveillance. Another possible surveillance measure can be derived by using persistence homology techniques like bar coding (40), a method of identifying topological invariants in cloud data. First, sequenced isolates can be transformed in high-dimensional space by using distance measures, such as pairwise genetic distance. From these cloud data, points can be linked together on the basis of certain criteria to form a simplicial complex, i.e., a network of points, line segments, triangles, and even n-dimensional counterparts. When this criterion is a distance less than some $\boldsymbol{\varepsilon}$, a filtered simplicial complex or Vietoris-Rips stream is created.

An objective in the study of topology is to identify features of filtered simplicial complexes that persist across all values of $\boldsymbol{\varepsilon}$. A useful characteristic is the Betti number, which in dimension 0 is b_0, the number of connected components in a particular simplicial complex. Trivially, with a large enough $\boldsymbol{\varepsilon}$, b_0 is 1, but what is interesting is the minimum value of $\boldsymbol{\varepsilon}$ that yields a b_0 value of 1. Low values would indicate a higher degree of surveillance.

The calculation of Vietoris-Rips streams for large point cloud data can be computationally expensive, and an alternative method called witness streams (45) can be used to estimate $\boldsymbol{\varepsilon}$. Suppose a landmark subset of points (L) is preselected from point cloud data either at random or using a max-min scheme. Let $d(L, p)$ be a vector of distances between a point p and all points in L. In dimension 0, a pair of points is then linked if there

exists a witness point z such that the maximum $d(L, p)$ is less than the sum of ε and the minimum $d(L, p)$.

Using the Javaplex software package at http://code.google.com/p/javaplex, we implemented a witness stream to filter different strains of influenza virus. Of N total sequences per strain, n sequences were chosen as landmarks by using the max-min selection algorithm, such that $N/n = 3$ (45). Bar coding was performed at filtration Hamming distances ε at intervals of 20 from 0 to 100.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00524-12/-/DCSupplemental.

Figure S1, PDF file, 0.4 MB.
Figure S2, PDF file, 0.1 MB.
Figure S3, PDF file, 0.2 MB.
Figure S4, PDF file, 0.1 MB.
Table S1, XLSX file, 0.1 MB.
Table S2, XLSX file, 0.1 MB.
Table S3, XLSX file, 0.1 MB.
Table S4, XLSX file, 0.1 MB.
Table S5, XLSX file, 0.1 MB.
Table S6, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

We thank the National Library of Medicine (R01 LM010140) and the National Institutes of Health (U54 CA121852-05) for funding.

REFERENCES

1. **WHO.** 2004. The world health Report 2004—changing history. World Health Organization, Geneva, Switzerland. http://www.who.int/whr/2004/en/.
2. **Morens DM, Folkers GK, Fauci AS.** 2004. The challenge of emerging and re-emerging infectious diseases. Nature **430**:242–249.
3. **Morse SS.** 1995. Factors in the emergence of infectious diseases. Emerg. Infect. Dis. **1**:7–15.
4. **Daszak P, Cunningham AA, Hyatt AD.** 2000. Emerging infectious diseases of wildlife—threats to biodiversity and human health. Science **287**:443–449.
5. **Stöhr K.** 2003. The global agenda on influenza surveillance and control. Vaccine **21**:1744–1748.
6. **Chan J, Holmes A, Rabadan R.** 2010. Network analysis of global influenza spread. PLoS Comput. Biol. **6**:e1001005. http://dx.doi.org/doi:10.1371/journal.pcbi.1001005.
7. **Butler D.** 2012. Flu surveillance lacking. Nature **483**:520–522.
8. **Webster RG, Govorkova EA.** 2006. H5N1 influenza—continuing evolution and spread. N Engl J. Med. **355**:2174–2177.
9. **Kawaoka Y.** 2012. H5N1: flu transmission work is urgent. Nature **482**:155.
10. **Fouchier RA, García-Sastre A, Kawaoka Y.** 2012. Pause on avian flu transmission studies. Nature **481**:443.
11. **Cohen J.** 2012. The limits of avian flu studies in ferrets. Science **335**:512–513.
12. **Le QM, Kiso M, Someya K, Sakai YT, Nguyen TH, Nguyen KH, Pham ND, Ngyen HH, Yamada S, Muramoto Y, Horimoto T, Takada A, Goto H, Suzuki T, Suzuki Y, Kawaoka Y.** 2005. Avian flu: isolation of drug-resistant H5N1 virus. Nature **437**:1108.
13. **Trifonov V, Khiabanian H, Rabadan R.** 2009. Geographic dependence, surveillance, and origins of the 2009 influenza A (H1N1) virus. N. Engl. J. Med. **361**:115–119.
14. **Ito T, Couceiro JN, Kelm S, Baum LG, Krauss S, Castrucci MR, Donatelli I, Kida H, Paulson JC, Webster RG, Kawaoka Y.** 1998. Molecular basis for the generation in pigs of influenza A viruses with pandemic potential. J. Virol. **72**:7367–7373.
15. **Kida H, Ito T, Yasuda J, Shimizu Y, Itakura C, Shortridge KF, Kawaoka Y, Webster RG.** 1994. Potential for transmission of avian influenza viruses to pigs. J. Gen. Virol. **75**:2183–2188.
16. **Scholtissek C.** 1990. Pigs as "mixing vessels" for the creation of new pandemic influenza A viruses. Med. Principles Pract. **2**:65–71.
17. **Layne SP, Monto AS, Taubenberger JK.** 2009. Pandemic influenza: an inconvenient mutation. Science **323**:1560–1561.
18. **Hernandez CX, Joseph C, Khiabanian H, Rabadan R.** 2011. Understanding the origins of a pandemic virus. arXiv:.4568. http://arxiv.org/ftp/arxiv/papers/1104/1104.4568.pdf.
19. **Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghwani J, Bhatt S, Peiris JS, Guan Y, Rambaut A.** 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature **459**:1122–1125.
20. **Bogner P, Capua I, Lipman DJ, Cox NJ, et al.** 2006. A global initiative on sharing avian flu data. Nature **442**:981.
21. **Karesh W.** 2011. Predict: surveillance and prediction for emerging pathogens of wildlife. BMC Proc **5**:1. http://dx.doi.org/doi:10.1111/j.1863-2378.2011.01439.x.
22. **Pavoine S, Ollier S, Pontier D.** 2005. Measuring diversity from dissimilarities with Rao's quadratic entropy: are any dissimilarities suitable? Theor. Popul. Biol. **67**:231–239.
23. **Cox NJ, Bender CA.** 1995. The molecular epidemiology of influenza viruses. Semin. Virol. **6**:359–370.
24. **Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC.** 2008. The genomic and epidemiological dynamics of human influenza A virus. Nature **453**:615–619.
25. **Araújo JMG, Nogueira RM, Schatzmayr HG, Zanotto PM, Bello G.** 2009 Phylogeography and evolutionary history of dengue virus type 3. Infect. Genet. Evol. **9**:716–725.
26. **Bertolotti L, Kitron U, Goldberg TL.** 2007. Diversity and evolution of west Nile virus in Illinois and the United States, 2002–2005. Virology **360**:143–149.
27. **WHO.** 2011 Influenza transmission zones. World Health Organization, Geneva, Switzerland. http://www.who.int/csr/disease/swineflu/Influenza_transmission_zones.pdf.
28. **USDA.** 2006. Release no. 0296.06. U.S. Department of Agriculture, Washington, DC. http://www.usda.gov/wps/portal/usda/usdahome?contentid=2006/08/0296.xml.
29. **Spackman E, Swayne DE, Suarez DL, Senne DA, Pedersen JC, Killian ML, Pasick J, Handel K, Pillai SP, Lee CW, Stallknecht D, Slemons R, Ip HS, Deliberto T.** 2007. Characterization of low-pathogenicity H5N1 avian influenza viruses from North America. J. Virol. **81**:11612–11619.
30. **Lanciotti RS, Lewis JG, Gubler DJ, Trent DW.** 1994. Molecular evolution and epidemiology of dengue-3 viruses. J. Gen. Virol. **75**:65–75.
31. **Chambers TJ, Halevy M, Nestorowicz A, Rice CM, Lustig S.** 1998. West Nile virus envelope proteins: nucleotide sequence analysis of strains differing in mouse neuroinvasiveness. J. Gen. Virol. **79**:2375–2380.
32. **Beatty ME, Stone A, Fitzsimons DW, Hanna JN, Lam SK, Vong S, Guzman MG, Mendez-Galvan JF, Halstead SB, Letson GW, Kuritsky J, Mahoney R, Margolis HS, Asia-Pacific and Americas Dengue Prevention Boards Surveillance Working Group.** 2010. Best practices in dengue surveillance: A report from the Asia-Pacific and Americas Dengue Prevention Boards. PLoS Negl. Trop. Dis. **4**:e890. http://dx.doi.org/doi:10.1371/journal.pntd.0000890.
33. **Eison M, Komar N, Sorhage F, Nelson R, Talbot T, Mostashari F, McLean R, West Nile Virus Avian Mortality Surveillance Group.** 2001. Crow deaths as a sentinel surveillance system for West Nile virus in the northeastern United States, 1999. Emerg. Infect. Dis. **7**:615–620.
34. **Andreadis TG, Anderson JF, Vossbrinck CR.** 2001. Mosquito surveillance for west Nile virus in Connecticut, 2000: isolation from *Culex pipiens*, *Cx. restuans*, *Cx. salinarius*, and *Culiseta melanura*. Emerg. Infect. Dis. **7**:670–674. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2631746/pdf/11585530.pdf.
35. **Pesko KN, Ebel GD.** 2012. West Nile virus population genetics and evolution. Infect. Genet. Evol. **12**:181–190.
36. **Gifford RJ, de Oliveira T, Rambaut A, Pybus OG, Dunn D, Vandamme AM, Kellam P, Pillay D, UK Collaborative Group on HIV Drug Resistance.** 2007. Phylogenetic surveillance of viral genetic diversity and the evolving molecular epidemiology of human immunodeficiency virus type 1. J. Virol. **81**:13050–13056.
37. **Yang Z, Rannala B.** 2012. Molecular phylogenetics: principles and practice. Nat. Rev. Genet. **13**:303–314.
38. **Drummond AJ, Suchard MA, Xie D, Rambaut A.** 2012. Bayesian phylogenetics BEAUti and the BEAST 1.7. Mol. Biol. Evol. **29**:1969–1973.
39. **Yang Z.** 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

40. **Carlsson G, Zomorodian A, Collins A, Guibas L.** 2004. Persistence bar codes for shapes, p. 124–135. Proceedings of the Eurographics/ACM SIGGRAPH Symposium on Geometry Processing. Association for Computing Machinery, Nice, France.
41. **Chao A, Colwell RK, Lin CW, Gotelli NJ.** 2009. Sufficient sampling for asymptotic minimum species richness estimators. Ecology **90**:1125–1133.
42. **Luoh SM, McGregor MW, Hinshaw VS.** 1992. Hemagglutinin mutations related to antigenic variation in H1 swine influenza viruses. J. Virol. **66**: 1066–1073.
43. **Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D.** 2008. The influenza virus resource at the National Center for Biotechnology Information. J. Virol. **82**:596–601.
44. **Miller GA.** 1955. Note on the bias of information estimates, p 95–100. *In* Information theory in psychology: problems and methods. Proceedings of a conference on the estimation of information flow, Monticello, Illinois, July 5-9, 1954, and related papers. Free Press, New York, NY.
45. **de Silva V, Carlsson G.** 2004. Topological estimation using witness complexes. Eurographics Symposium on Point-Based Graphics. Eurographics, Geneva, Switzerland. http://comptop.stanford.edu/u/preprints/witness.pdf.