LETTERS

Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution

Elodie Ghedin[1], Naomi A. Sengamalay[1], Martin Shumway[1], Jennifer Zaborsky[1], Tamara Feldblyum[1], Vik Subbu[1], David J. Spiro[1], Jeff Sitz[1], Hean Koo[1], Pavel Bolotov[2], Dmitry Dernovoy[2], Tatiana Tatusova[2], Yiming Bao[2], Kirsten St George[3], Jill Taylor[3], David J. Lipman[2], Claire M. Fraser[1], Jeffery K. Taubenberger[4] & Steven L. Salzberg[1,5]

Influenza viruses are remarkably adept at surviving in the human population over a long timescale. The human influenza A virus continues to thrive even among populations with widespread access to vaccines, and continues to be a major cause of morbidity and mortality[1,2]. The virus mutates from year to year, making the existing vaccines ineffective on a regular basis, and requiring that new strains be chosen for a new vaccine. Less-frequent major changes, known as antigenic shift, create new strains against which the human population has little protective immunity, thereby causing worldwide pandemics. The most recent pandemics include the 1918 'Spanish' flu, one of the most deadly outbreaks in recorded history, which killed 30–50 million people worldwide, the 1957 'Asian' flu, and the 1968 'Hong Kong' flu[3]. Motivated by the need for a better understanding of influenza evolution, we have developed flexible protocols that make it possible to apply large-scale sequencing techniques to the highly variable influenza genome. Here we report the results of sequencing 209 complete genomes of the human influenza A virus, encompassing a total of 2,821,103 nucleotides. In addition to increasing markedly the number of publicly available, complete influenza virus genomes, we have discovered several anomalies in these first 209 genomes that demonstrate the dynamic nature of influenza transmission and evolution. This new, large-scale sequencing effort promises to provide a more comprehensive picture of the evolution of influenza viruses and of their pattern of transmission through human and animal populations. All data from this project are being deposited, without delay, in public archives.

The genomes reported here comprise the initial results from the Influenza Genome Sequencing Project, a partnership between the US National Institute of Allergy and Infectious Diseases and collaborators from around the world[4], whose goal is to sequence the genomes of thousands of influenza virus isolates. This study is the first, to our knowledge, to attempt to sequence strains that were not pre-selected for particular virulence or other unusual characteristics, and should therefore provide a relatively unbiased view of influenza virus strains in the population. Here we focus on a collection from New York State spanning several years, and subsequent studies will focus on samples from multiple, distant geographical sources across a longer time span. As our analysis shows, even within a geographically constrained set of isolates, we have found surprising genetic diversity, indicating that the reservoir of influenza A strains in the human population—and the concomitant potential for segment exchange between strains—may be greater than was previously suspected.

The genome of the influenza A virus (family Orthomyxoviridae) consists of eight single-stranded negative sense RNA molecules spanning approximately 13.5 kilobases (kb). The segments range in length from 890 to 2,341 nucleotides and encode a total of 11 proteins. Although a large number of partial influenza A virus sequences now exist in the public archives (for example, GenBank), relatively few complete genomes are available. In part this is due to the technical difficulty of constructing an efficient sequencing pipeline for an RNA-based organism. The bulk of the public data on influenza comprises short fragments from the haemagglutinin (HA) or neuraminidase (NA) segments of the genome, which encode the two main surface proteins and which, it is widely believed, are the source of most of the antigenic variation in the virus. As a result of this project, the number of complete human H3N2 influenza virus genomes in GenBank has already grown from just seven genomes to over 200.

We have completely sequenced all eight segments from 207 H3N2 isolates and two H1N2 isolates. In total, the finished sequence covers 2,821,103 bases, with an average of 13,498 bases per isolate. Table 1 shows the sequencing results for all isolates, broken down by segment. The polymerase chain reaction with reverse transcription (RT–PCR)-based sequencing strategy produces an average of 5.6 sequencing reads covering each nucleotide, as shown in Table 1. The average Phred quality value[5] was 33, which at 5.6-fold coverage corresponds to an error rate of 3.2×10^{-19}; however, regions of low coverage will have higher error rates. These regions can be inspected at the NCBI Assembly Archive[6], which displays the raw data underlying every nucleotide in each genome. Note that the error rate in these genomes is likely to be considerably lower than previously sequenced influenza isolates, which in many cases reflect single-pass or two-pass sequencing. Assembly was performed using the Minimus assembler program[7] followed by the AutoEditor program to correct erroneous bases[8]. Details of assembly and annotation are provided in the Supplementary Methods.

This is the first large-scale analysis of influenza isolates collected in a relatively unbiased manner, allowing a comprehensive look at an influenza virus population across several seasons within a constrained geographical area. Among RNA viruses, only human immunodeficiency virus (HIV) has been subjected to similar whole-genome analysis[9,10]. In this first set of 209 genomes, we have observed multiple, novel mutational events, including point mutations, deletions and segment exchange. By carefully cataloguing these events, we can begin to get the first real picture of the rate of mutational events underlying influenza A virus evolution.

[1]The Institute for Genomic Research, 9712 Medical Center Dr., Rockville, Maryland 20850, USA. [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA. [3]Wadsworth Center, New York State Department of Health, Albany, New York 12201, USA. [4]Department of Molecular Pathology, Armed Forces Institute of Pathology, Rockville, Maryland 20850, USA. [5]Center for Bioinformatics and Computational Biology, University of Maryland Institute for Advanced Computer Studies, College Park, Maryland 20742, USA.

Although extensive previous research has catalogued changes in the HA and NA segments, we found new mutations in these segments as well as multiple changes in the other six segments. Some of these changes are shown in Fig. 1 and discussed below, and a comprehensive list of amino acid mutations in all eight segments is provided in the Supplementary Data.

Perhaps the most dramatic finding in our data is the discovery of an epidemiologically significant reassortment that explains the appearance, during the 2003–2004 season, of the 'Fujian/411/2002'-like strain, for which the existing vaccine had limited effectiveness. In a recent paper[11], we described how phylogenetic analysis of 156 H3N2 genomes from our project revealed the clear presence of multiple, distinct clades circulating in the population. Through a reassortment event, a minor clade provided the haemagglutinin gene that later became part of the dominant strain after the 2002–2003 season. Two of our samples, A/New York/269/2003 (H3N2) and A/New York/32/2003 (H3N2), show that this minor clade continued to circulate in the 2003–2004 season, when most other isolates were reassortants. In these samples (Fig. 1, see columns for November 2003) the HA segment is clearly similar to the dominant clade, whereas the other segments all show numerous differences.

This finding illustrates not only that the influenza virus population contains multiple lineages at any given time, but also that alternate, minor lineages can contribute genetic variation to the dominant lineage, resulting in epidemiologically significant, antigenically novel strains. It is worth emphasizing that our sequence-based sampling approach—in contrast to traditional serologically based sampling—will reveal co-circulating strains even before they become antigenically novel.

Figure 1 illustrates five seasons' worth of mutations in all proteins from the 207 H3N2 influenza virus isolates included in this study. For clarity, amino acid positions are shown only if they underwent genetic changes in at least three isolates. Each mutation is indicated by a colour shift along a row in the figure. For example, the first row shows that the amino acid in position 5 of HA1 mutated from glycine (G, shown in light green) to valine (V, shown in burgundy) in November 1999, and then back to glycine in November 2001 and afterwards, except for three isolates in January to February 2002 that show a glutamic acid (E, shown in pink) at that position. In total, 186 positions experienced at least one amino acid change.

As the figure shows, mutations appear both during and between influenza seasons. For example, HA residues 5, 33 and 92 remained unchanged from May 1999 to October 1999, and then mutated in November 1999, leading to a permanent switch for the rest of that season. Notably, multiple changes in the internal segments, including those encoding the polymerase genes (PA, PB1 and PB2), the nucleocapsid protein (NP), and two non-structural proteins (NS1 and NS2), first appeared in the 2001–2002 season and became fixed thereafter.

Data from isolates collected in the spring of 2003 provide a glimpse of the transitional period before a major reassortment event. Many of the HA mutations that became dominant during the 2003–2004 influenza season first appeared in February 2003. Mutations to residues 155 and 156 of the HA1 domain (H155T and Q156H) show up in early 2003; these sites are accessible to antibodies and had an important role in the antigenic mismatch between the vaccine strain and the circulating viruses in the 2003–2004 season[12]. A different picture emerges in the other proteins, where mutations that appear in February 2003 remain only in a few isolates. This clearly indicates that a reassortment event brought in a new HA segment during or before the spring of 2003, and subsequent data show that this reassortant strain became dominant in the 2003–2004 influenza season[11].

A number of important mutations found in our data may affect receptor-binding affinity and potentially increase viral replication efficiency. Studies have determined that changes in HA residues 183, 186 and 226 could affect HA receptor-binding affinity[13], and residue positions 131, 222, 225 and 226 are important for efficient replication[12]. Mutation S186G appears in circulating viruses during the 2001–2002 influenza season, along with mutation V202I, and remains in the 2003–2004 season. Mutations A131T, W222R and G225D also emerge in February 2003. The HA1 T155H and H156Q mutations in our data are accompanied by a possibly correlated mutation at residue 25 (L25I).

The neuraminidase protein has a box-shaped globular head with four catalytic sites that allow the cleavage of sialic acid linkages[14]. Amino acid positions important for antigenic drift have been identified for the N2 subtype[14] as well as other regions likely to be involved in virus–host interactions and qualified as phylogenetically important regions[15]. Sequence data in our study indicate that once residue 197—an antigenic site[14,16]—mutated from 197H to 197D early in the 1999–2000 influenza season, it was accompanied by the mutation R249K. This residue is probably not in a functional site but may be functionally compensating by maintaining the accessibility of surface residues. Residue 199 interestingly switched (E199K) for the 2003–2004 influenza season for the majority of the isolates, except for the two isolates corresponding to the minor non-reassorted clade (A/New York/269/2003 and A/New York/32/2003)[11].

Table 2 lists correlated mutations that may be co-mutations; that is, where there appears to be a balancing effect between two sites on the same protein, or between a site on an internal protein and one on a surface protein. The best example is seen for T392 in NA, which is present in the same eight isolates (appearing in 2001–2003) where there is an I463 mutation in PB2 (see Fig. 1).

The fact that the minor Fujian-like clade has donated its HA to the previously dominant strain rather than itself becoming the dominant circulating virus indicates that there may be important amino acid co-substitutions in the other proteins essential for viral fitness[11]. When comparing the NA and internal proteins of the dominant circulating major clade present during the 2003–2004 influenza season with the previous dominant clade, there are a few

Table 1 | Sequencing results for 209 complete genomes of H3N2 and H1N2 human influenza A viruses

Segment	Length (nt)	Size of coding region (nt)	Total finished sequence (nt)	Finished sequence per segment (nt)	Coding completeness (%)	Average segment coverage*
PB1	2,341	2,274	483,932	2,315	100	5.5
PB2	2,341	2,280	482,950	2,311	100	5.3
PA	2,233	2,151	463,810	2,219	100	5.9
HA	1,762	1,701	364,012	1,742	100	5.8
NP	1,565	1,497	324,681	1,553	99.9	5.4
NA	1,466	1,410	303,410	1,452	99.7	5.9
M	1,027	982	213,806	1,023	100	5.2
NS	890	838	184,502	883	100	5.7
Total	13,625	13,133	2,821,103	13,498	99.95	5.6

nt, nucleotide.
*Coverage is defined as the average number of individual sequencing reads covering each nucleotide of finished sequence.

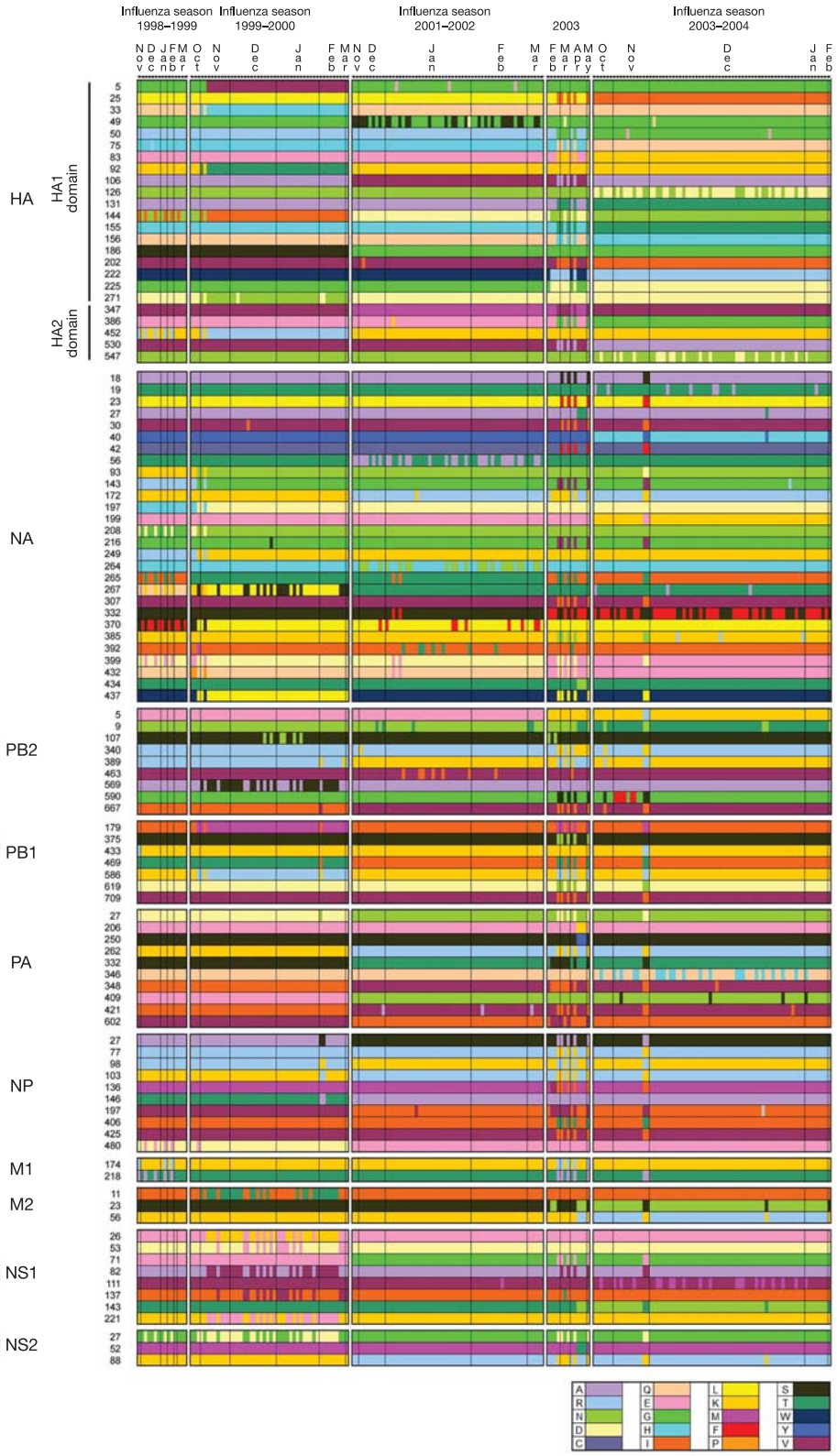

Figure 1 | Sites with genetic changes across the ten main proteins in 207 influenza A viruses. Each row represents a single amino acid position in one protein. Amino acids (single-letter abbreviations are used) are colour-coded as shown in the key, so that mutations can be seen as changes in colour when scanning from left to right along a row. For simplicity, only amino acids that showed changes in at least three isolates are shown. Each column represents a single isolate, and columns are only a few pixels wide in order to display all 207 H3N2 isolates in this figure. Isolates are ordered along the columns chronologically according to the date of collection; boundaries between influenza seasons are indicated by gaps between columns. A more detailed version of this figure, showing positions that experienced any amino acid change and showing identifiers for the isolates in each column, is available as Supplementary Fig. 1.

Table 2 | Correlated mutations within and between influenza virus proteins

Correlated mutations within segments						
HA*	NA	PB2	PB1	PA	NP	NS1
N30–S400	L45–D401	R389–I667	I179–K586	D27–K262	A27–R98	Q25–N127
S30–T400	P45–G401	K389–V667	M179–R586	N27–R262	S27–K98	K25–D127
K269–V476	H197–R249		D619–V709	E206–S250	1406–V425	E26–K221
R269–A476	D197–K249		N619–I709	K206–Y250	T406–I425	K26–E221
D275–L304					A451–S482	
I144–R252					T451–N482	
					I136–I425	

Correlated mutations of surface and internal protein residues			
HA*–NA	HA*–M1	HA*–PA	NA–PB2
G49–T56	F159–K208	G275–N359	T392–I463
S49–A56		D275–D359	
N547–D147			
D547–N147			

*HA residues are counted from the first amino acid of the HA1 domain.

substitutions that appear to be unique. NA Y40H and E199K, for example, occur in regions potentially affecting virus–host interactions[15]. On the basis of the NA alignments performed using data from the LANL influenza database (data not shown), the E199K mutation is also seen in southeast Asian isolates that were collected during the 2003–2004 influenza season when the Fujian-like variant was the dominant circulating virus in that region. Over the rest of the NA protein, however, the southeast Asian isolates resemble the non-reassortant North American clade B[11]. This indicates that this clade may have had deficiencies in certain residues that affected its ability to become the dominant virus. There is, unfortunately, no publicly available data for these proteins from the southeast Asian isolates with which to do a comparative analysis to the North American major and minor clades.

Recent reports[17–19] have described a newly discovered protein, known as PB1-F2, encoded by a shifted reading frame in the PB1 gene. The data presented here more than double the total number of complete PB1 segments in the public archives, and we found that the PB1-F2 open reading frame is preserved in 206 out of 209 PB1 genes. In most cases (180 out of 206) the protein's length is 90 amino acids, but it is 87 amino acids in 23 isolates and 80 amino acids in three isolates. In three cases, an in-frame stop truncates the predicted protein after 11 amino acids. The translations of PB1-F2 for all 209 isolates are provided in the Supplementary Data.

Our project includes two clear examples of segment exchange between H1N1 and H3N2 viruses, both of which are H1N2 serotypes. In both isolates from our collection, only the haemagglutinin segment was exchanged, and these appear to be descendants of a human–swine recombinant, as has been reported previously[20]. Although segment exchange has been reported before[21,22], no accurate data on the frequency of these events have been collected. Our observation of three events (two exchanges between different H3N2 clades[11], and one exchange between H3N2 and H1N1) in 209 samples may provide an initial baseline for future estimates.

The Influenza Genome Sequencing Project is currently being expanded to include avian influenza, in an effort to establish how often these strains cross the species barrier and move into the human population. One possible cause for influenza pandemics is the mixing through reassortment of an avian influenza strain with a human strain via co-infection of a single host[23]. Recent reports of transmissions of avian influenza virus to humans[24,25] have raised concerns that a new pandemic might emerge[26]. Despite the importance of the threat that influenza presents, no previous effort has been made to study its complete genome on a large scale. The protocols described here are being generalized to include large numbers of avian influenza isolates that, like the genomes reported here, will be deposited immediately in public archives.

METHODS

All sequence data used in this study are available from GenBank, and also via a project page at http://www.tigr.org/flu. In addition, all 209 genomes and GenBank accession numbers are available as a single file in the Supplementary Data.

All samples for this study were collected by the Virus Reference and Surveillance Laboratory of the Wadsworth Center in Albany, New York, which maintains a repository of human influenza samples dating back to 1992. Virus samples were received as part of outbreak investigations, through the reference function of the laboratory, and, since 2001, as part of a sentinel physician influenza programme. Use of the diagnostic samples in this study was approved by the New York State Department of Health Institutional Review Board.

Viral RNA isolation. Isolates were amplified in tube cultures of primary rhesus monkey kidney (pRhMK) cells before extracting 140 μl of culture supernatant. Viral RNA was extracted from clarified supernatant fluid using the Qiagen BioRobot M48 workstation with the MagAttract Viral RNA M48 kit (Qiagen).

RNA ligation. RNA was circularized overnight at 4 °C with T4 RNA ligase (Epicentre). Before the ligation step, the RNA was first treated with tobacco acid pyrophosphatase (20 U TAP in a 15-μl reaction, incubated at 37 °C for 1 h). TAP treatment is usually used to remove molecules from the 5′ end of RNA, mostly plus-strand RNA. Although no such molecules are expected to be present on the influenza genomic RNA segments, ligation was more efficient with this treatment than without. The circularized RNA was cleaned again with the RNeasy Mini kit (Qiagen).

RT–PCR and sequencing. The first step in the high-throughput sequencing pipeline uses reverse transcription followed by polymerase chain reaction amplification to generate overlapping DNA amplicons covering each segment of the influenza virus genome. Overlapping primers were designed approximately every 200–250 nucleotides along the genome; degenerate primers allow the pipeline to tolerate sequence variation. In order to capture the extreme ends of each segment, we used an RNA circularization step before the RT–PCR[27]. We then used RT–PCR to amplify a chimaeric product that contained the sequence from both ends of the segment.

Complementary DNA synthesis. RT–PCRs were performed with a OneStep RT–PCR kit (Qiagen). Ninety-five reactions were performed per RNA sample. Degenerate primers were designed based upon the alignment of selected human H3N2 sequences. For most of the segments, all full-length and nearly full-length sequences from 1980 to the present were aligned and used for primer design. For others, more stringent criteria were used in order to reduce the number of sequences in the set to a more manageable number. An M13 sequence tag was added to the 5′ end of each primer to be used for sequencing (F primers: TGTAAAACGACGGCCAGT; R primers: CAGGAAACAGCTATGACC). Eight pairs of primers were designed to span the ligated ends of each segment to capture the end sequences. Four of the reactions were analysed on an agarose gel for quality control purposes. Primer sequences are included as a separate table in the Supplementary Data.

Amplicons were prepared for sequencing by incubating them at 37 °C for 60 min with 0.5 U of shrimp alkaline phosphatase (Amersham) and 1 U of exonuclease I (Amersham) to inactivate remaining dNTPs and to digest the single-stranded primers. The enzymes were inactivated by incubation at 72 °C for 15 min.

Sequencing. Sequencing reactions were performed on a standard high-throughput sequencing system using Big Dye Terminator chemistry (Applied Biosystems) with 2 μl of template cDNA. Each amplicon was sequenced from each end using M13 primers (F primer: TGTAAAACGACGGCCAGT; R primer: CAGGAAACAGCTATGACC). Sequencing reactions were analysed on an Applied Biosystems 3730 ABI sequencer. Each influenza isolate was processed on its own 96-well plate to minimize the possibility of sample mix-ups.

Data release. Raw traces were submitted to the NCBI Trace Archive. The finished assembly of each isolate, showing how the traces are aligned to one another and to the finished sequence, was deposited in the NCBI Assembly Archive[6], which allows scientists to investigate the data supporting every nucleotide of each genome. An annotation pipeline developed at NCBI (see Supplementary Methods) was run to make gene assignments, and finished genomes with annotation were deposited without delay in GenBank.

Received 7 July; accepted 16 September 2005.
Published online 5 October 2005.

1. Simonsen, L., Fukuda, K., Schonberger, L. B. & Cox, N. J. The impact of influenza epidemics on hospitalizations. *J. Infect. Dis.* **181,** 831–837 (2000).
2. Thompson, W. W. *et al.* Influenza-associated hospitalizations in the United States. *J. Am. Med. Assoc.* **292,** 1333–1340 (2004).
3. Cox, N. J. & Subbarao, K. Global epidemiology of influenza: past and present. *Annu. Rev. Med.* **51,** 407–421 (2000).
4. Fauci, A. S. Race against time. *Nature* **435,** 423–424 (2005).
5. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
6. Salzberg, S. L., Church, D., DiCuccio, M., Yaschenko, E. & Ostell, J. The genome Assembly Archive: a new public resource. *PLoS Biol.* **2,** E285 (2004).
7. Pop, M., Phillippy, A., Delcher, A. L. & Salzberg, S. L. Comparative genome assembly. *Brief. Bioinform.* **5,** 237–248 (2004).
8. Gajer, P., Schatz, M. & Salzberg, S. L. Automated correction of genome sequence errors. *Nucleic Acids Res.* **32,** 562–569 (2004).
9. Harris, M. E. *et al.* Among 46 near full length HIV type 1 genome sequences from Rakai District, Uganda, subtype D and AD recombinants predominate. *AIDS Res. Hum. Retroviruses* **18,** 1281–1290 (2002).
10. Mikhail, M. *et al.* Full-length HIV type 1 genome analysis showing evidence for HIV type 1 transmission from a nonprogressor to two recipients who progressed to AIDS. *AIDS Res. Hum. Retroviruses* **21,** 575–579 (2005).
11. Holmes, E. C. *et al.* Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol.* **3,** e300 (2005).
12. Jin, H. *et al.* Two residues in the hemagglutinin of A/Fujian/411/02-like influenza viruses are responsible for antigenic drift from A/Panama/2007/99. *Virology* **336,** 113–119 (2005).
13. Lu, B., Zhou, H., Ye, D., Kemble, G. & Jin, H. Improvement of influenza A/Fujian/411/02 (H3N2) virus growth in embryonated chicken eggs by balancing the hemagglutinin and neuraminidase activities, using reverse genetics. *J. Virol.* **79,** 6763–6771 (2005).
14. Colman, P. M., Varghese, J. N. & Laver, W. G. Structure of the catalytic and antigenic sites in influenza virus neuraminidase. *Nature* **303,** 41–44 (1983).
15. Fanning, T. G., Reid, A. H. & Taubenberger, J. K. Influenza A virus neuraminidase: regions of the protein potentially involved in virus-host interactions. *Virology* **276,** 417–423 (2000).
16. Air, G. M., Els, M. C., Brown, L. E., Laver, W. G. & Webster, R. G. Location of antigenic sites on the three-dimensional structure of the influenza N2 virus neuraminidase. *Virology* **145,** 237–248 (1985).
17. Malide, D., Yewdell, J. W., Bennink, J. R. & Cushman, S. W. The export of major histocompatibility complex class I molecules from the endoplasmic reticulum of rat brown adipose cells is acutely stimulated by insulin. *Mol. Biol. Cell* **12,** 101–114 (2001).
18. Gibbs, J. S., Malide, D., Hornung, F., Bennink, J. R. & Yewdell, J. W. The influenza A virus PB1–F2 protein targets the inner mitochondrial membrane via a predicted basic amphipathic helix that disrupts mitochondrial function. *J. Virol.* **77,** 7214–7224 (2003).
19. Chanturiya, A. N. *et al.* PB1–F2, an influenza A virus-encoded proapoptotic mitochondrial protein, creates variably sized pores in planar lipid membranes. *J. Virol.* **78,** 6304–6312 (2004).
20. Marozin, S. *et al.* Antigenic and genetic diversity among swine influenza A H1N1 and H1N2 viruses in Europe. *J. Gen. Virol.* **83,** 735–745 (2002).
21. Xu, X. *et al.* Reassortment and evolution of current human influenza A and B viruses. *Virus Res.* **103,** 55–60 (2004).
22. Lindstrom, S. E., Cox, N. J. & Klimov, A. Genetic analysis of human H2N2 and early H3N2 influenza viruses, 1957–1972: evidence for genetic divergence and multiple reassortment events. *Virology* **328,** 101–119 (2004).
23. Reid, A. H., Taubenberger, J. K. & Fanning, T. G. Evidence of an absence: the genetic origins of the 1918 pandemic influenza virus. *Nature Rev. Microbiol.* **2,** 909–914 (2004).
24. Cyranoski, D. Vaccine sought as bird flu infects humans. *Nature* **422,** 6 (2003).
25. Abbott, A. & Pearson, H. Fear of human pandemic grows as bird flu sweeps through Asia. *Nature* **427,** 472–473 (2004).
26. Li, K. S. *et al.* Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature* **430,** 209–213 (2004).
27. Szymkowiak, C. *et al.* Rapid method for the characterization of 3′ and 5′ UTRs of influenza viruses. *J. Virol. Methods* **107,** 15–20 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors acknowledge the laboratory and bioinformatics assistance provided by D. Kosack, C. Hauser, L. Groveman, R. Halpin, A. Phillippy and J. Sparenborg at TIGR, and S. Griesemer, M. Kleabonas and R. Bennett at the Wadsworth Center. We also thank M. Giovanni for help with project oversight and direction. This work was supported in part by the US National Institute of Allergy and Infectious Diseases. Viruses described in this study include some isolates collected as part of the Sentinel Physician Influenza Surveillance Program, which is supported by the US Centers for Disease Control and Prevention.

Author Information GenBank accession numbers are provided in the Supplementary Data. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.L.S. (salzberg@umd.edu).