THE ROYAL SOCIETY

doi 10.1098/rstb.2001.0998

The haemagglutinin gene, but not the neuraminidase gene, of 'Spanish flu' was a recombinant

Mark J. Gibbs[*], **John S. Armstrong** and **Adrian J. Gibbs**

School of Botany and Zoology, Faculty of Science, Australian National University, Australian Capital Territory 0200, Australia

Published analyses of the sequences of three genes from the 1918 Spanish influenza virus have cast doubt on the theory that it came from birds immediately before the pandemic. They showed that the virus was of the H1N1 subtype lineage but more closely related to mammal-infecting strains than any known bird-infecting strain. They provided no evidence that the virus originated by gene reassortment nor that the virus was the direct ancestor of the two lineages of H1N1 viruses currently found in mammals; one that mostly infects human beings, the other pigs. The unusual virulence of the virus and why it produced a pandemic have remained unsolved. We have reanalysed the sequences of the three 1918 genes and found conflicting patterns of relatedness in all three. Various tests showed that the patterns in its haemagglutinin (HA) gene were produced by true recombination between two different parental HA H1 subtype genes, but that the conflicting patterns in its neuraminidase and non-structural–nuclear export proteins genes resulted from selection. The recombination event that produced the 1918 HA gene probably coincided with the start of the pandemic, and may have triggered it.

Keywords: orthomyxovirus; 1918 flu; recombination; selection; haemagglutinin; neuraminidase

> ... nothing remarkable was seen in the HA or the NA of H1N1-1918.
>
> (Lederberg 2001)

1. INTRODUCTION

The need to understand the origin and virulence of the virus that caused the 1918 pandemic is plain. In *America's forgotten pandemic: the influenza of 1918*, historian Alfred W. Crosby (1989) records in graphic detail the 'Spanish flu' pandemic. It was the most severe outbreak of acute human disease ever recorded. The virus first appeared in late 1917 or, more certainly, early 1918, killed 20 million or more people worldwide within a few months and then, apparently, disappeared. Commenting on its severity Crosby noted, 'Influenza and pneumonia, when they kill, usually kill those two extremes of life, the very young and the old' (p. 21), whereas, for the Spanish flu, 'an extraordinary high proportion of the dead were young adults. Nearly one-half of the dead were between 20 and 45 years of age.... No other influenza before or since has had such a propensity for pneumonic complications. And pneumonia kills' (p. 27).

Since the nature of viruses was first properly recognized just over one century ago, it has become clear that genetic changes in the viruses themselves often trigger disease outbreaks. Unlike all cellular organisms, most viruses have RNA genomes. None of the replicases of such viruses has been found to have a proofreading function and, probably as a result, their mutation rate is often more than a million-fold faster than that of cellular organisms. This great capacity for change is constrained by selection for functionality and fitness. For many years, it was also thought that the evolution of RNA viruses was also limited because their genomes were unable to recombine like those encoded in DNA. Viruses with new combinations of genomic segments produced by reassortment (pseudo-recombination) were detected in the 1950s, but true recombinants with chimeric genes or genomes copied from parts of several progenitors were not detected until more recently. However, it has gradually become clear, largely through bioinformatic analysis of gene sequences, that recombination is a major driving force in the evolution of RNA viruses, especially those with positive (messenger-sense) genomes, although there is less experimental evidence that those with negative-sense genomes, like influenza viruses, recombine to produce new chimeric genomic molecules, and the fact that negative-sense genomes recombine is still not accepted by some.

Very little was known about the cause of influenza when the 1918 pandemic occurred. It is now known that the primary worldwide ecological niche of influenza A viruses is the digestive tract of water birds; most variants of this very diverse group of viruses have been isolated from ducks and gulls (Webster *et al.* 1992), and in those hosts the virus usually causes few or no symptoms. All influenza viruses found in mammals seem to have originated from water birds, and mammal-infecting strains can acquire genes from this reservoir. This was shown to be the trigger for the 1957 and 1968 pandemics as the novel influenza isolates then found in the human population had virions that were very closely related serologically to

[*] Author for correspondence (mark.gibbs@anu.edu.au).

influenzas from birds. When, later, the genes of these isolates were sequenced it was found that those of the haemagglutinin (HA), neuraminidase (NA) and RNA polymerase (PB1) were replaced in the 1957 pandemic strain and the HA and PB1 genes were replaced in the 1968 pandemic strain, and the other genes were unchanged.

The changes could only have occurred in mixed infections of bird-infecting and human-infecting strains that allowed reassortment to generate the pandemic strains with new combinations of the eight genomic negative-sense RNA segments. The HA and NA proteins of influenza viruses form spikes that cover the outer lipid membrane of the virions of influenza viruses; the HA is multifunctional and involved in the infection of susceptible cells, whereas the NA is involved in the egress of progeny virions from infected cells. The virulence of influenza A viruses is determined by several factors, some encoded by their HAs; minor mutational changes in the HA gene have produced highly pathogenic forms of the virus and are probably important in permitting bird-infecting strains to adapt to mammals (Gambaryan *et al.* 1997; Khatchikian *et al.* 1989; Klenk & Garten 1994; Subbarao *et al.* 1998). Antibodies against the HA neutralize the infectivity of influenza viruses and the NA is also antigenic, so when the HA and NA genes are exchanged 'antigenic shifts' occur, and the new reassorted strains are able to side-step immunity of the human population acquired by prior infection. The antigenic shifts that occurred with the 1957 and 1968 strains allowed them to spread widely and rapidly and produce major pandemics. Thus, questions about virulence and origins became linked.

In view of the 1957 and 1968 pandemics, discussions of the likely cause of the 1918 pandemic have centred on the possibility that the 1918 virus also emerged from birds or that at least that some of its genes came from an avian strain (Gorman *et al.* 1991; Webster 1999). Among the first influenza viruses to be isolated and cultured was the 'classical swine' isolate A/Sw/Iowa/30 (Iowa 30). Tests with antibodies from the survivors of the 1918 pandemic showed a close serological relationship with Iowa 30, indicating that the 1918 virus was probably an influenza A virus of the H1N1 subtype, which was the dominant subtype in the human population for most of the 20th century. However, interpretation of the serological relationship was, and sometimes still is, overstated. Iowa 30 has been shown to be related to all other subtype A influenzas, and serology is no longer considered the best way to distinguish relationships between closely related virus strains. The picture became more complex when it was realized that there were three main lineages of H1N1 viruses, those that were isolated mostly from humans, mostly from pigs and mostly from birds, and it was not known which of these produced the 1918 virus.

When the gene sequences from several isolates of the two main mammal lineages of the H1N1 subtype were compared their differences were found to correlate broadly with differences in their time of isolation, so that the date of divergence of the lineages could be estimated by regression. It was assumed that the divergence point of the mammalian lineages marked the date at which the common ancestral H1N1 virus first switched from birds into mammals. Depending on the gene sequences that were compared, estimates of this emergence date ranged from 1900 to 1918, and although most regressions indicated that it occurred several years before 1918, the approximate concordance with the 1918 pandemic was taken as support for the theory that the 1918 virus emerged immediately before the pandemic and that the two main lineages of influenza A in mammals were its descendants (Gorman *et al.* 1991). Any discrepancies were easily discounted as the analyses assumed that there was a constant rate of change, which clearly does not occur; on average, influenza sequences change at a rate of 0.5–1.0% per year, but much faster rates have been calculated from comparisons of some pandemic strains and those that have recently switched hosts. Moreover, selection of the sequences used for the analyses may have introduced bias.

Confirmation seemed to be at hand when, in a spectacular technical feat, the first sequences from the fragments of the genomic RNA of the 1918 virus were obtained from preserved tissues of three of its victims, two who died in the eastern USA in September 1918 and the third in Alaska a month later. The complete gene sequences of the HA and NA and also the bi-cistronic gene of the non-structural–nuclear export proteins (NS/NEP) have so far been reported (Basler *et al.* 2001; Reid *et al.* 1999, 2000; Taubenberger *et al.* 1997). These sequences have confirmed that the pandemic was caused by an influenza A of the H1N1 subtype, but when compared with the homologous genes of related viruses they yielded few clues about how the pandemic was initiated nor why it was so severe (Basler *et al.* 2001; Lederberg 2001; Martindale 2000; Reid *et al.* 1999, 2000; Taubenberger *et al.* 1997, 2000; Webster 1999).

Careful phylogenetic analysis did not yield quite the predicted results, as they showed that the 1918 virus was most closely related to other H1N1 influenza isolates from mammals and was only distantly related to isolates from birds. Reid *et al.* (1999) concluded that 'the 1918 HA gene has accumulated enough non-selected mammalian associated changes to place it consistently in the mammalian clade' (p. 1655). This was contrasted with the analyses of the HA and NA genes from the 1957 and 1968 pandemic strains, which were both placed in avian clades and had much closer bird-infecting relatives. The phylogenetic analyses of the 1918 genes were also consistent with the regression analyses; Reid *et al.* (1999) suggested that 'the pandemic virus had been adapting in mammals before 1918' (p. 1656) and Reid *et al.* (2000) concluded 'it is possible that these genes (HA and NA) have been adapting in mammalian hosts for some time before emerging in pandemic form' (p. 6790). It was not clear which kind of mammal originally hosted the progenitors of the 1918 virus, as the 1918 sequences were placed on a basal branch in the phylogenetic trees, sometimes joined to the swine lineage and sometimes joined to the human lineage, depending on which method of analysis was used and which component of the sequences. Therefore, far from resolving the issues, the results were inconclusive and puzzling. The possibility that an additional factor or factors may have triggered the pandemic began to be considered (Basler *et al.* 2001; Lederberg 2001).

Most analyses of influenza gene sequences, including the initial analyses of the 1918 sequences, have involved

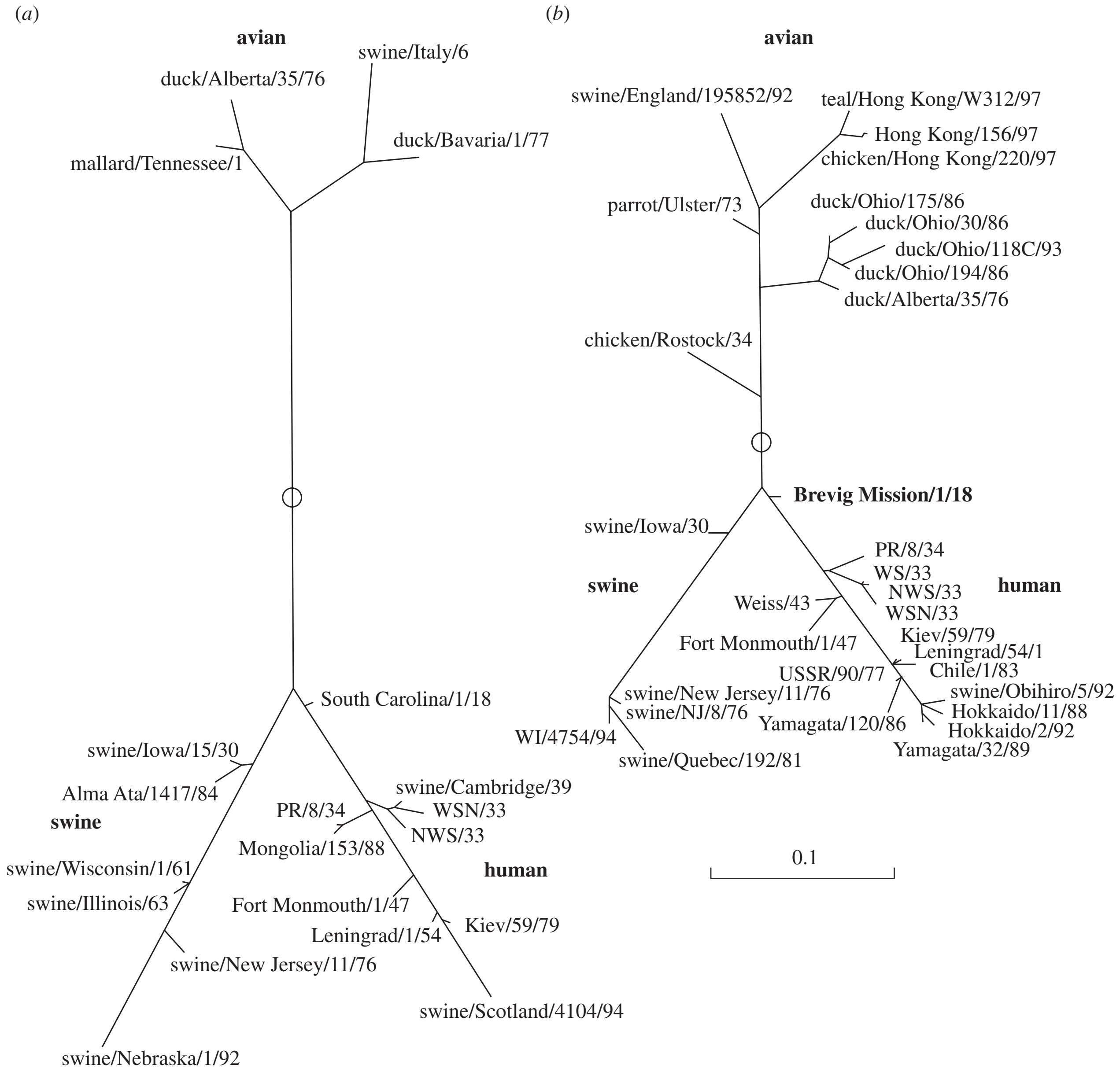

Figure 1. Phylogenetic trees calculated from the complete (*a*) HA and (*b*) NA gene sequences by the ML quartet puzzling method (Strimmer & Von Haeseler 1996) and using the Tamura & Nei (1993) model of substitution with eight γ-rate categories. ML branch lengths are shown and tree midpoints are marked with open circles. The code names of the isolates and the GenBank accession codes (HA and/or NA) are: A/South Carolina/1/18 (AF117241), A/swine/Nebraska/1/92 (S67220), A/swine/New Jersey/11/76 (K00992, AF250363), A/swine/Illinois/63 (X57493), A/swine/Wisconsin/1/61 (AF091307), A/swine/Iowa/15/30 (AF091308, AF250364), A/Alma Ata/1417/84 (S62154), A/swine/Cambridge/39 (D00837), A/WSN/33 (J02176, L25817), A/NWS/33 (U08903, L25815), A/Mongolia/153/88 (Z54287), A/PR/8/34 (NC.002017, NC002018), A/Fort Monmouth/1/47 (U02464, AF250357), A/Leningrad/54/1 (M38309, M38312), A/Kiev/59/79 (M38353, M38335), A/USSR/90/77 (K02018), A/swine/Scotland/4104/94 (AF085413), A/chicken/Hong Kong/220/97 (AF046081), A/Hong Kong/156/97 (AF046089), A/teal/Hong Kong/W312/97 (AF250481), A/parrot/Ulster/73 (K02252), A/chicken/Rostock/34 (X52226), A/Chile/1/83 (X15281), A/Yamagata/120/86 (D31948), A/Hokkaido/2/92 (D31945), A/swine/Obihiro/5/92 (D31947), A/Yamagata/32/89 (D31950), A/Hokkaido/11/88 (D31944), A/WS/33 (L25816), A/swine/Quebec/192/81 (U86144), A/New Jersey/8/76 (M27970), A/Wisconsin/4754/94 (U53166), A/duck/Ohio/175/86 (AF250358), A/duck/Ohio/30/86 (AF250359), A/duck/Ohio/194/86 (AF250360), A/duck/Alberta/35/76 (AF250362), A/duck/Ohio/118C/93 (AF250361), A/swine/England/195852/92 (AF250366), A/Weiss/43 (AF250365), A/Brevig Mission/1/18 (AF250356).

comparisons of complete gene sequences. Gene phylogenies have been mapped from these comparisons and reassortments have been detected (Subbarao *et al.* 1998; Kilbourne 1977; Webster *et al.* 1992). However, smaller-scale anomalies of intragenic relatedness, such as those resulting from recombination or selection, can only be clearly detected when subsequences of genes, or particular components of genes (e.g. synonymous and non-synonymous nucleotide changes) are analysed and compared. In this paper, we report analyses of the HA and

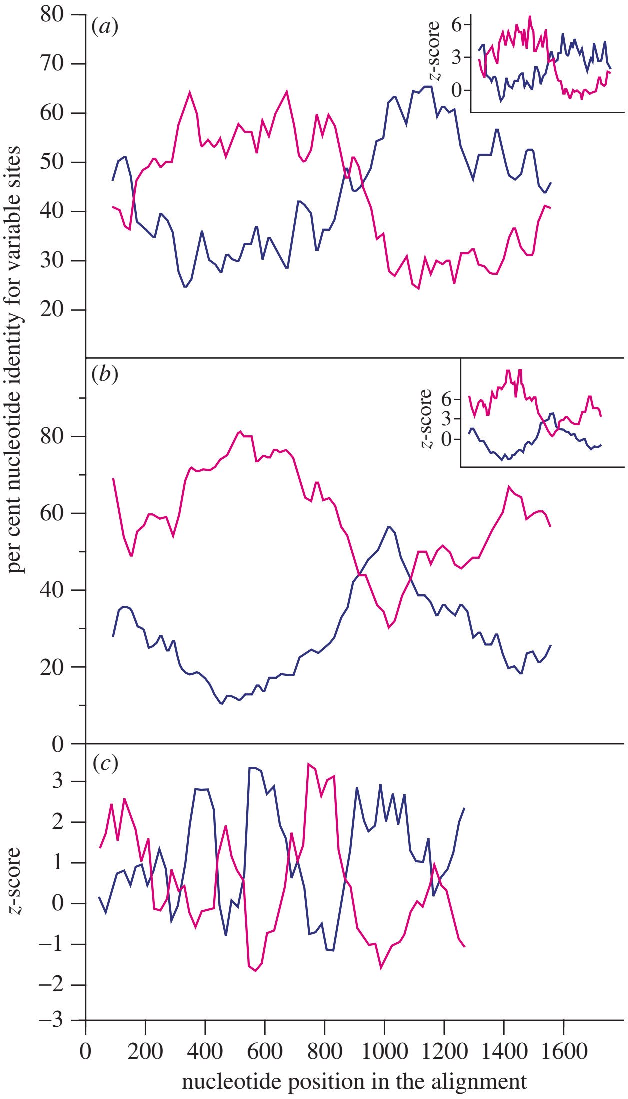

Figure 2. CPRs in different regions of HA and NA genes determined by the sister-scanning method (Gibbs *et al*. 2000). (*a*) Per cent nucleotide identity scores (inset: *z*-values) for all variable sites in the HA gene of the 1918 influenza virus calculated from an alignment with the HA genes of the variants A/swine/Wisconsin/1/61 (red plots) and A/Kiev/59/79 (blue plots). (*b*) Per cent nucleotide identity scores (inset: *z*-values) for all variable sites in the HA gene of the classical swine isolate A/swine/Iowa/15/30 calculated from an alignment with the HA genes of the variants A/swine/Wisconsin/1/61 (red plots) and A/Kiev/59/79 (blue plots). (*c*) Mean *z*-scores of 10 comparisons with the Brevig 1918 NA gene sequence. Two reference sequences were used in each comparison: one from the avian N1NA gene lineage and one from human N1NA gene lineage. The mean Brevig/avian relatedness is shown by the blue line, and the mean Brevig/human relatedness by the red line.

NA gene sequences of the 1918 influenza using, primarily, the sister-scanning method to test for recombination. These have revealed that the 1918 HA gene, but not the NA gene, was a recombinant, and that it was produced by a recombination event that probably coincided with the start of the 1918 pandemic. We suggest that this recombination event, rather than reassortment or host switching was the genetic trigger for the 1918 pandemic (Gibbs *et al.* 2001*a*) and summarize the facts in this paper.

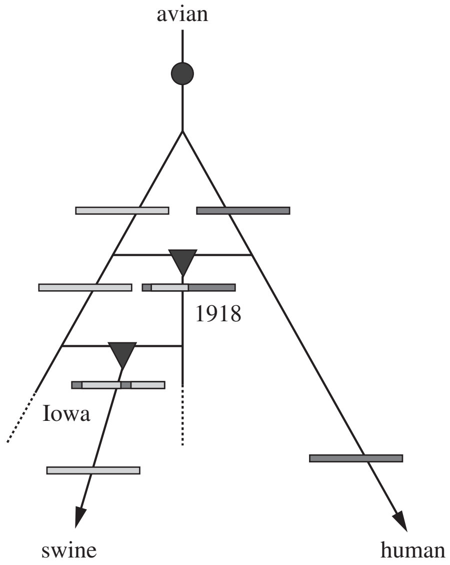

Figure 3. A cartoon mapping the likely evolution of different parts of the HA genes of the 1918 and Iowa influenza viruses, and the two main lineages, human and swine, of the H1 subtype influenza HA gene. A common ancestor of the viruses switched hosts (circle) from birds to mammals sometime before 1918. Two recombination events (triangles) linked the early lineages but some of these lineages died out (dotted lines).

2. RECOMBINATION DETECTION

The sister-scanning program SiScan 2.0 (available free from http://online.anu.edu.au/bozo/software/index.html; Gibbs *et al.* 2000, 2001*b*) compares three aligned sequences using, as an outlier, a fourth sequence generated by local randomization of one or two of the real sequences. A window is passed over the four aligned sequences and, at each window position, all the possible patterns of identity are counted and recorded. A Monte Carlo randomization is then done to the four nucleotides in each column of the alignment, position by position within the current window. *z*-scores are calculated for each pattern of identity separately, or they are summed in various ways, so as to assess the relatedness of pairs of sequences. In this way, the significance of the patterns detected in the four sequences is tested against the background of all the other patterns in the alignment, and the dominant one detected; when one signal is significant, those for the others are usually much lower and are often close to zero. Nucleotide identity scores for pairs of sequences are calculated from the patterns, but the method recognizes and tests patterns, and is not a distance-matrix method relying on pairwise comparisons.

Although in most comparisons the relatedness of the sequences are similar in all parts of the sequences, others may show that one pair of sequences is significantly related in some parts of the sequence, but that another pair is significantly related in other parts. These 'conflicting patterns of relatedness' (CPRs) are possible evidence for recombination or for differential selection, where one part of the sequence changes more quickly than others as a result of selection. The SiScan 2.0 program also permits the patterns of relatedness to be assessed using sites of a particular type. For example 'informative sites', as defined for parsimony analyses, can

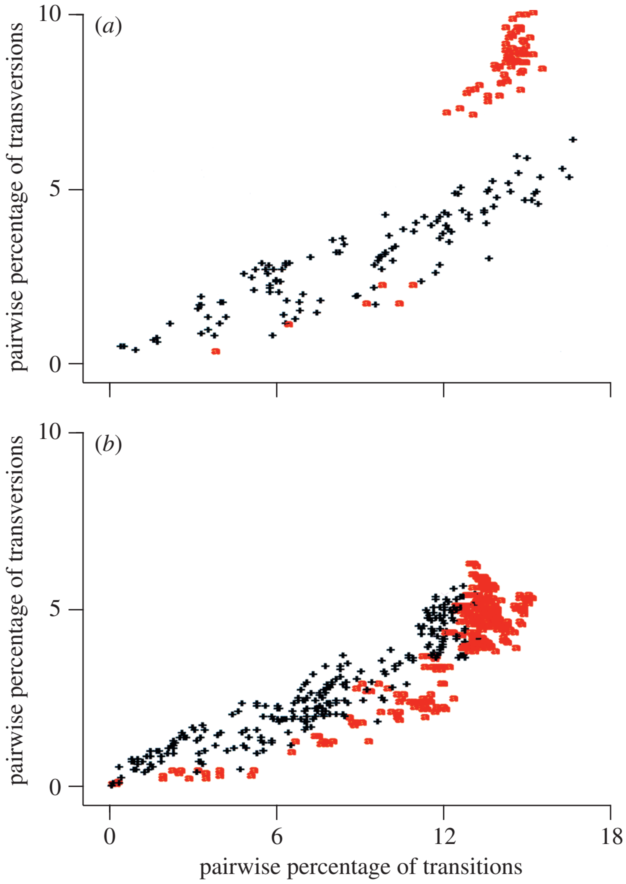

Figure 4. Graphs showing the percentage of transversion differences plotted against the percentage of transition differences for pairs of (*a*) HA gene sequences and (*b*) NA gene sequences. Graphs were made using the DIPLOMO program (Weiller & Gibbs 1993, 1995). Points calculated from pairwise comparisons involving an avian sequence are marked in red and all others in black.

be selected, as too can the first or second or third codon positions, or sites that differ synonymously or non-synonymously.

We used the 'likelihood analysis of recombination in DNA' (LARD) method (Holmes *et al.* 1999) as the main means to confirm our SISCAN results, as its maximum-likelihood (ML) approach is completely different from that of sister-scanning, and it has been used before to detect and confirm evidence of recombination in the sequences of other viruses (Worobey & Holmes 2001; Worobey *et al.* 1999), as well as in bacterial and human genes and pseudogenes.

3. HA GENE SEQUENCES

We compared the complete sequences of the HA genes of 30 H1 subtype isolates. They were selected from the international databases, aligned by CLUSTALX (Jeanmougin *et al.* 1998) using default parameters, and gaps representing two codons were removed. The resulting alignment was 1695 nucleotides long; its 1026 5′-terminal nucleotides encoded the HA1 and the remainder the HA2. Phylogenetic trees (figures 1 and 5) were inferred by a method based on ML (tree-puzzle) (Strimmer & Von Haeseler 1996), or by a distance method (neighbour joining (NJ)) (Saitou & Nei 1987). When trees were inferred from the complete sequences, by whichever method, they confirmed that the sequences clustered into one or other of three main lineages; those that had been isolated mostly, but not exclusively, from human beings, from pigs or from birds. The HA gene of the 1918 influenza (A/South Carolina/1/18) fell close to the node linking the three main lineages, whichever method of phylogenetic analysis was used.

SISCAN analyses detected statistically significant CPRs in the HAs of the 1918 influenza and those of the Iowa ('classical swine') lineage, namely the serologically closely related isolates A/swine/Iowa/15/30 (Iowa/30), A/Alma Ata/1417/84 (Alma Ata) and A/swine/St-Hyacinthe/148/90 (St-Hyacinthe) (Beklemishev *et al.* 1993; Bikour *et al.* 1995). The CPRs were detected in the four HA genes using several independent combinations of reference sequences, and this confirmed their reliability. The analyses showed that the 1918 sequence contained a region from around nucleotides 120–775 that was most closely related to the homologous regions of swine-lineage genes, with regions on either side that were most closely related to human-lineage sequences (figure 2*a*).

The SISCAN analyses also indicated that the HA1 coding region of all three Iowa lineage viruses were probably inherited directly from the 1918 HA gene, as they had the same pattern of CPRs as the 1918 HA in their HA1 gene; however, their HA2 region was related to swine-lineage HA2s not human-lineage HA2s (figure 2*c*).

Four out of five of the likely recombination sites between the CPRs were also found using LARD, and shown to be statistically significant (Holmes *et al.* 1999). SISCAN analyses were also done using, separately, the nucleotide positions with differences that were either 'synonymous' or 'non-synonymous', or only from third codon positions, or only from 'informative sites', as described in parsimony analyses, and all produced similar significant CPRs. Importantly, the synonymous differences yielded statistically significant evidence of CPRs (figure 2*b*) only when the 1918 HA sequence was compared with pairs of reference sequences, one of which was a human-lineage sequence and the other a swine-lineage sequence, but never when one of the pair was an avian-lineage sequence. The simplest explanation of these results is that the 1918 sequence was a recombinant of two distinct HA sequence lineages that pre-dated the 1918 HA gene (figure 3).

The structure of an H3HA has been determined (Wiley & Skehel 1987). The H1HA and H3HA genes and their encoded proteins are unequivocally homologous, so it is clear that the part of the H1HA gene between nucleotides 120 and 775 encodes the globular domain of the HA, whereas the other parts of the gene encode the N- and C-terminal parts of the HA1 which, together with all the HA2 polypeptide, form the stalk that anchors the HA to the lipid outer layer of the virion. Thus, the part of the 1918 HA gene most closely related to present swine-lineage HA genes encodes its globular domain, which includes its major antigenic sites, the host-cell–receptor binding site and almost all the glycosylation sites, which are also important to its antigenicity, whereas its stalk domain is from a virus related to present human-lineage

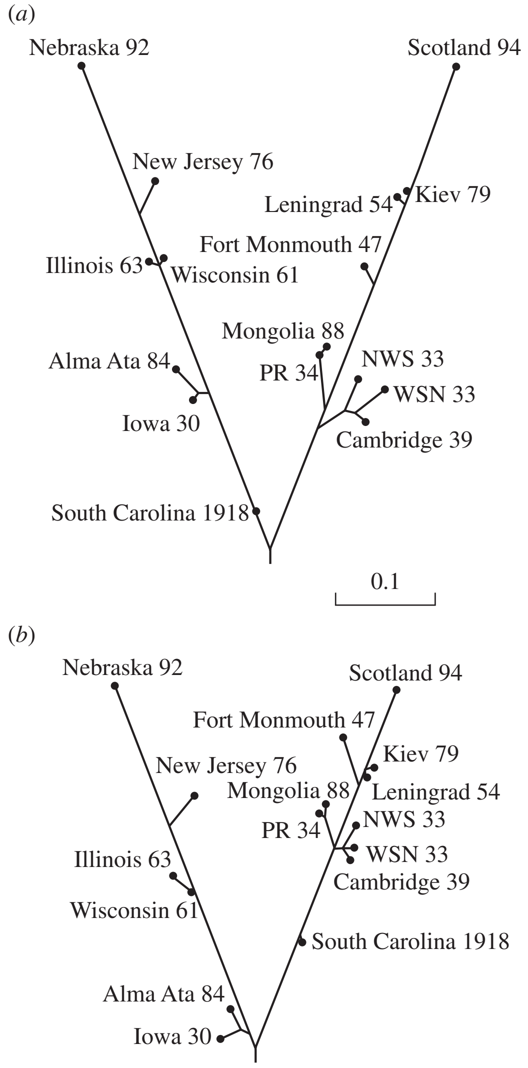

Figure 5. Phylogenies of the nucleotide sequences encoding most of the HA1 and all the HA2 domain of mammalian H1 influenzas. ML trees were calculated from aligned nucleotide positions 310–870 (*a*) and 1070–1650 (*b*) using the same method that was used for the complete gene trees (figure 1). Both trees are rooted at the midpoint between the terminal nodes for Iowa 30 and NWS 33. The dates of isolation of the isolates are shown next to taxon names. The code names of the isolates and the GenBank accession codes for their HA genes are as in figure 1.

viruses. However, although the HA1s of Iowa-lineage HAs are closely related to the HA1 of the 1918 HA gene and have closely similar CPRs, they have a swine-lineage HA2, so their stalk region is a molecular chimera of swine-lineage and human-lineage parts.

The simplest scenario that would explain the SiScan results (figure 3) is that an avian influenza H1HA gene became established in mammals some time before 1918 and diverged into two lineages that had their own distinct patterns of nucleotide differences. Those two lineages recombined in a mixed infection to produce the 1918 gene. The Iowa-lineage HA gene was produced by a further recombination event that replaced the HA2 region of the 1918 HA gene with a swine-lineage HA2 region.

Phylogenetic trees calculated from the complete HA gene sequences by the ML method using a range of parameters, or by the NJ method, differed considerably in the positions and lengths of the branches connecting the avian sequences to the mammalian sequences, and also differed in the position of the 1918 sequence. This result was partly explained when, for each pair of sequences, transitional differences (i.e. purine to purine, or pyrimidine to pyrimidine changes) were plotted against transversional differences (i.e. purine to pyrimidine or the reverse), and it became clear (figure 4*a*) that avian and mammalian H1HAs were not members of the same population. Similar differences were found when pairwise 'synonymous' differences were plotted against 'non-synonymous' differences, as had been reported earlier (Gibbs *et al.* 1982). These results indicate that avian sequences are not appropriate outliers for locating the root of the mammalian H1HA tree in phylogenetic analyses, despite their widespread use for this purpose.

ML trees were therefore calculated from the mammalian isolate sequences alone and separate phylogenetic trees were calculated for two major regions of the sequences that did not include recombination sites, and which the SiScan analyses indicated had different histories. One region (nucleotide positions 300–870) encompassed the region encoding the HA globular domain, and the other (nucleotide positions 1100–1650) that encoding most of the stalk domain. The ML trees obtained from these two regions were congruent with those obtained from the complete sequences (figure 1*a*), in that the sequences formed two major lineages (figure 5*a*,*b*): the swine and human lineages. In the globular domain tree (figure 5*a*), the 1918 partial sequence was always on the swine-lineage side of the tree, regardless of how the root was estimated (e.g. midpoint of the nearest, most distant or average sequence), and in the HA2 tree it was always on the human-lineage side of the tree. These results were consistent with our conclusion that the 1918 HA gene was a recombinant.

In the globular domain tree (figure 5*a*), the 1918 sequence was placed on the 'trunk' connecting the human and swine lineages, whereas in the HA2 tree it was joined to the trunk by a branch so short that the HA2-encoding sequence of the 1918 virus probably differed from the predicted ancestral (trunk) sequence at only two or three sites (0.4%). Thus, both regions of the 1918 HA gene had probably changed very little since it was generated by recombination, and this suggests that the time between the origin of the 1918 virus and its 'preservation' was, at most, 1 year (Gibbs *et al.* 2001*a*). The victims from whom the 1918 influenza RNAs were obtained died in the major 'second wave' of the pandemic in late September and October 1918 (Crosby 1989; Reid *et al.* 1999); therefore, the 1918 HA gene was probably generated in late 1917 or early 1918. The earliest record of the pandemic in North America was the 'first wave' in early 1918, which, although less lethal than the 'second wave' in late 1918, was probably caused by the same virus with environmental or host factors causing differences in morbidity, and there is one report suggesting that an outbreak occurred in Europe in late 1917 (J. S. Oxford, personal communication). This close coincidence of the start of the pandemic and a recombination event that could produce the genetic novelty required to trigger it suggests that one may have caused the other.

The ML trees also provided further detail of the other early post-1918 events indicated by SiScan analysis. There is no evidence of recombination in the human-lineage HA genes so it seems likely that they did not evolve from the 1918 gene but, instead, from one of the parents of the 1918 HA gene. Both parts of the sequences of the Iowa HA lineage, which, as mentioned, probably originated by recombination, are directly linked to later swine isolates, which also appear not to be recombinants. It is unclear whether the human-lineage 'stalk' regions of the Iowa HA gene were lost by recombination or whether the human-lineage signal was lost by selection; the HA1 sequences encoding part of the stalk are too short, and the differences too small, for this to be resolved.

4. NA AND NS/NEP GENE SEQUENCES

The complete NA gene sequences of 32 H1 subtype isolates, including that of the 1918 virus, A/Brevig Mission/1/18 (Brevig) (Reid *et al.* 2000), were selected from the international databases, and aligned by ClustalX (Jeanmougin *et al.* 1998) using default parameters. Trees calculated by NJ and ML (figure 1*b*) methods confirmed that the NA sequences clustered into one or other of three main lineages that closely correlated with those of the HA sequences described above; those mostly from birds, from pigs or from human beings. Scatter plots of the transitional and transversional differences between the NA gene sequences showed (figure 4*b*) that, as with the H1HAs, avian and mammalian H1NAs are separate populations, evolving in different modes, and similar evidence was found by plotting pairwise synonymous against non-synonymous differences. So a simple phylogenetic tree, like that in figure 1*b*, calculated from all nucleotide positions in the sequences and containing differences resulting from several causes is unlikely to depict accurately the phylogeny of the N1NA gene.

A SiScan analysis showed that six of the NA sequences, including Brevig, showed statistically significant CPRs. To simplify interpretation of the results, all of these, except Brevig, were removed from the dataset, and the remainder reanalysed using all sequence positions or, separately, only those that, in each comparison of three sequences and a random outlier, were 'synonymous' or were 'non-synonymous'. The analyses using 'synonymous' differences alone gave no significant CPRs with the Brevig NA sequence, and so only those found using 'non-synonymous' differences were examined in more detail.

Sixty-nine pairs of sequences gave statistically significant CPRs with the Brevig sequence when 'non-synonymous' differences were analysed. In 67 of these, one of each pair was an avian-lineage sequence and the other from a mammal, 62 of them human. The fact that most of these CPRs were obtained from nucleotides shared with the sequences of avian isolates, the ancestral lineage, and with human isolates, the lineage of descendants, indicates that they resulted from the differential selection of parts of the encoded NA protein, most probably after the switch from birds to mammals.

Plots of either the 'raw scores', or the z-values, of the comparisons giving significant CPRs were closely similar, and showed two major and two minor regions of affinity with human-lineage sequences, interspersed with three major regions of affinity with avian sequences. The peaks of these regions were consistent in position (±25 nucleotides) and phase, but differed in magnitude in the range $z = 2.5$–5.0, so the mean z-values were calculated from the 10 comparisons that included the largest z-values (figure 2*d*). X-ray diffraction analysis of one N2 sub-type NA has shown it to consist of 'six topologically identical β-sheets arranged in a propeller formation' (Varghese *et al.* 1983, p. 35) and, as with the 1918 HA, it was clear that the pattern of affinities in the gene sequence correlated with the structure of the encoded protein.

So the amino-acid sequences of the Brevig NA and the five human and five bird reference NAs that gave the clearest CPRs were aligned, and it was found that 40 out of the 388 amino-acid residues of the Brevig NA were the same as those of avian isolate NAs, and 18 were the same as those in human isolates. These were plotted onto the homologous positions of the structure of the NA 'head', and it can be seen that although the avian-specific sites are dispersed throughout the structure, the human-specific sites are clustered around the side of the NA subunit that is closest to the outer membrane of the virion but away from the four-fold axis of the NA tetramer.

This distribution of altered sites is quite different from that resulting from antigenic drift, which usually involves the known antigen epitopes (Colman *et al.* 1983; Webster *et al.* 1982) that cover the surface of the NA subunit around the catalytic site, and hence are distal to the outer membrane of the virion. Interestingly, the distribution of avian–human NA differences on the virion-proximal surface of the NA subunit is closely similar to the pattern that was found when N9 NAs from a tern isolate and from a whale isolate were compared (Air *et al.* 1987). Fanning *et al.* (2000) used parsimony methods to identify 'phylogenetically important regions' (PIRs) in the amino-acid sequences of N1 NAs. Five of the 12 PIRs they identified are among the 11 regions that delimit the altered regions we have identified.

The NS/NEP gene is the only other gene from the 1918 influenza virus reported so far (Basler *et al.* 2001). We compared this sequence in SiScan analyses with the sequences of 38 other NS/NEP genes and found that it produced significant CPRs when compared using non-synonymous sites but not synonymous sites; 27 pairs of the sequences gave CPRs with the 1918 NS sequence and 26 of those pairs consisted of one avian-lineage sequence and one human sequence. Thus, the NS/NEP gene shows evidence of the same type of selection as the 1918 NA gene.

5. CRITICISMS

In this paper, and the preceding paper on the 1918 HA gene (Gibbs *et al.* 2001*a*), we have presented the simplest explanation of the CPRs found in the three genes of the 1918 virus, although, of course, other more complex and less likely interpretations are possible. Our conclusions are totally consistent with the facts, interpretations and speculation presented in several reviews (e.g. Taubenberger *et al.* 2000; Webster 1999). However, our conclusions have been criticized in at least six significant ways.

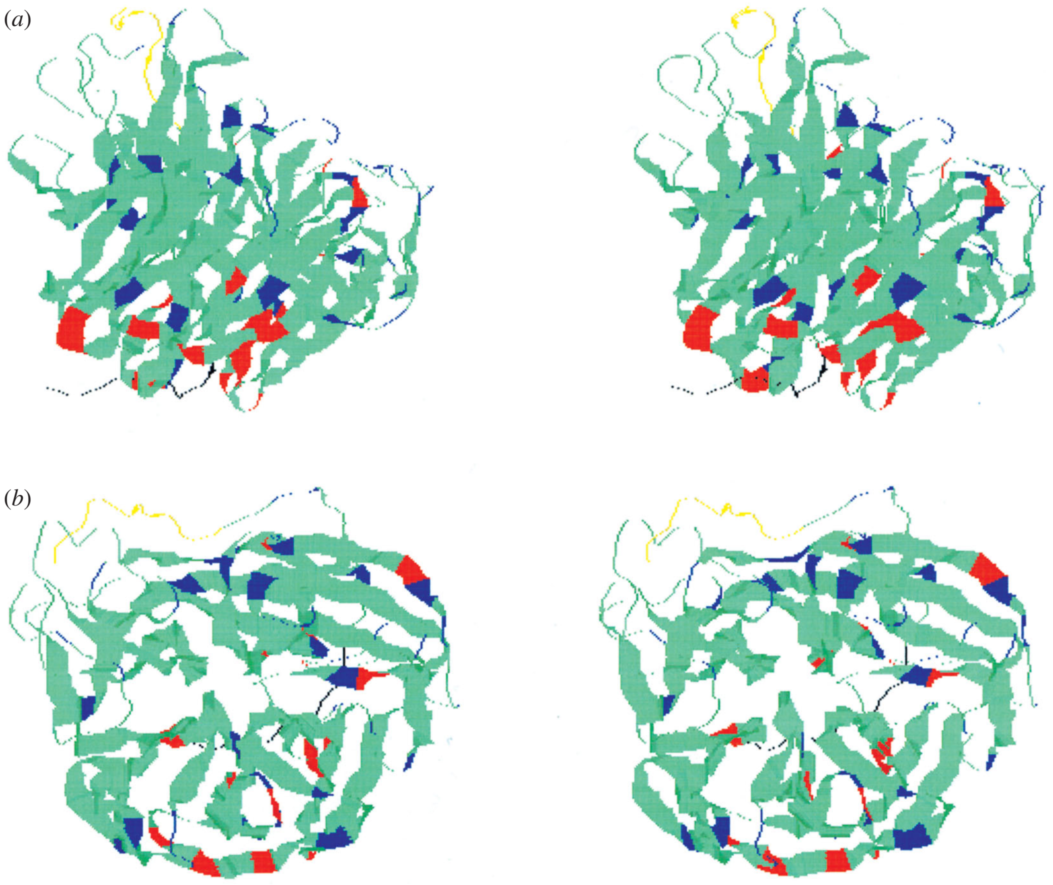

Figure 6. Stereo ribbon cartoons of the structure of the influenza NA subunit showing the position of residues that are shared with the proteins of most avian isolates but not human (blue), or with most human isolates but not avian (red). The NA subunit is seen (*a*) side view with the axis of the tetrameric protein at right angles to the plane of the page and top left or (*b*) top view looking into the active site of the enzyme with the axis of the protein in the plane of the page and behind on the left of the subunit. The N-terminal 10 residues of the subunit are coloured black, and the C-terminal 10 residues yellow.

(a) *Criticism 1. Recombination in the HA sequence occurred in the laboratory during polymerase chain reaction amplification or sequencing*

This can be discounted, as three independent samples of the 1918 genomic RNAs yielded near identical sequences (Reid *et al.* 1999) that included the recombination sites. Furthermore, the Alma Ata, Iowa and St-Hyacinthe HA genes, which were obtained and sequenced entirely independently from isolates obtained more than 10 years after the pandemic, have CPRs in the same regions. This, together with the results of phylogenetic analyses (figure 5), suggests not only that this early swine lineage inherited its HA1 region from the 1918 virus, but also that our interpretation of the CPRs in the 1918 HA sequence is correct.

(b) *Criticism 2. Homologous recombination has not been demonstrated experimentally between influenza viruses, so our results are speculative*

There is already experimental evidence of recombination in influenza A viruses. Influenza A viruses have been found with recombinant HA genes containing foreign sequences (Khatchikian *et al.* 1989; Röhm *et al.* 1996), defective interfering RNAs produced by recombination have also been found (Duhaut & Dimmock 1998; Fields & Winter 1982) and recombination has been detected in experimental systems (Bergmann *et al.* 1992; Morgan & Dimmock 1992; Orlich *et al.* 1994). Both the defective interfering RNAs and the experimentally generated recombinants (Fields & Winter 1982; Morgan & Dimmock 1992) were homologous recombinants in the most commonly used sense (Lai 1992). Bioinformatic analysis has already proven an invaluable and accurate method for detecting recombination in viruses; recombination was first detected in human immune deficiency virus by bioinformatic analysis (Robertson *et al.* 1995) and only later confirmed experimentally.

(c) *Criticism 3. Regions of the HA gene of the 1918 virus are more similar to those of later swine-lineage viruses than those of human-lineage viruses*

It has been suggested that regions of the HA gene of the 1918 virus are more similar to those of later swine-lineage viruses than those of human-lineage viruses because the 1918 virus resembles the common ancestor of both the swine and human lineages, but that whereas the human-lineage genes have changed dramatically, the swine-lineage genes have changed little (J. K. Taubenberger, quoted in Pickrell (2001)).

The H1 subtype HA swine-lineage sequences are not evolving significantly more slowly than the human-lineage genes; the branch lengths of the swine and human lineages in both the globular domain and HA2 trees are closely similar, as also shown in earlier reports (Reid *et al.* 1999). Branch-lengths are estimates of the difference between the sequences, so the trees indicate that the swine sequences have changed as much as the human ones. The same result was obtained when differences were calculated by directly comparing sequences independently of the trees. It is true that the antigenic sites change more rapidly in the human-lineage HAs than in swine-lineage HA, but these sites constitute only a small fraction of the globular domain. Thus, an argument based only on the antigenic sites cannot explain the consistent relationship across the whole sequence that encodes the globular domain. Furthermore, evidence of recombination was also found when the selected nucleotide sites, including those that code for antigenic sites, were removed from the analysis, as was done when only synonymous sites were analysed. The converse of this criticism, that the HA stalk-encoding parts of the gene changed rapidly to adapt to a human host, but that the globular domain-encoding sequence has evolved more slowly, can be discounted on similar grounds; the proposal cannot explain the large scale of the pattern and there is no evidence of unusually rapid change in the HA stalk domain of any influenza isolate, which, in all, has changed more slowly than the globular domain.

(d) *Criticism 4. Swine-like sites and human-like sites are mixed across the 1918 HA gene sequence*

This is true, but the two kinds of sites are unevenly distributed on a large scale and hence recombination is the most likely explanation. There is also an obvious reason for the sites being mixed: influenza A viruses are mutating at about 0.5–1% of nucleotide sites per year, and in the more recent sequences that were used to detect relatedness patterns in the 1918 sequences, point mutations have overwritten the original pattern. At the time of the recombination event, the swine and human lineages were about 3–5% different; by the 1990s they were more than 30% different.

(e) *Criticism 5. Recombination cannot be proven without showing a change in the branching order of a phylogenetic tree*

There is no requirement to use phylogenetic trees to show that recombination has occurred (Khatchikian *et al.* 1989; Maynard Smith & Smith 1998; McGuire *et al.* 1997; Sawyer 1989; Weiller 1998; Zhang *et al.* 2001). However, both of the methods we mostly used, sister-scanning and LARD, test the support for simple trees of either three or four taxa (Gibbs *et al.* 2000; Holmes *et al.* 1999).

(f) *Criticism 6. There is no evidence that pigs were infected with influenza viruses before 1918*

It may be true that there is no record of influenza-like symptoms in pigs before 1918. However, pigs are genetically diverse and have been domesticated worldwide for many millennia (Huang *et al.* 1999; Signer *et al.* 2000), and more subtypes and strains of influenza, mostly avian, have been isolated from swine than from human beings since 1918 (Olsen *et al.* 2000). It is therefore most likely that pigs were infected with influenzas before 1918, but that this was just neither recognized nor recorded.

6. EPILOGUE

Previous comparisons of the 1918 HA and NA genes with those of other H1 N1 isolates have not revealed why the virus was so virulent. Some components of the virulence of influenzas are determined by their HAs. It is likely that the two pre-1918 mammalian H1 HAs differed in at least 15 of the amino acids in their globular domains when they recombined to form the 1918 HA (Gibbs *et al.* 2001*a*). These differences may have directly, or indirectly via conformational changes, altered the antigenicity of the HA, so that those who had survived earlier infections were susceptible to the recombinant virus. Alternatively, or simultaneously, the receptor-binding, membrane-fusion or other functions of the HA protein may have been altered. Historical records show that most of the Spanish flu's victims were in their 20s, the prime of life, and many died of acute viral pneumonia before secondary infections occurred. In other influenza epidemics, most infections are of the upper respiratory tract and most victims among the very young or old. Thus, the virulence of the Spanish flu may have resulted, in part, from the recombinant HA providing the virus with a novel tissue specificity that enabled it to infect lungs; the elevated morbidity of the 20-year-olds may have occurred because their immune systems reacted more intensely than those of other age groups, and it was this intense reaction that produced the lethal pneumonia from which they died (R. V. Blanden, personal communication).

We have no idea why the avian–human NA differences are clustered on the virion-proximal surface of the NA subunit, unlike most of those acquired during antigenic drift or under antibody selection in experiments. Air *et al.* (1987) were also surprised when they found the same distribution of differences between N9 NAs from a whale isolate and a tern isolate. Both comparisons involve an avian-to-mammal switch and only experiments will distinguish between the possibility that the changes are selected by the switch in environment, from gut to respiratory tract, or differences between the cell membranes of birds and mammals; however, there is no obvious pattern in the types of amino acid (i.e. size, charge or polarity) that have changed during these two NA avian–mammal host switches.

We thank Charmaine Simeonovic, Georg Weiller and the Division of Botany and Zoology, Australian National University for supporting this unfunded research, and Paul Gibbs for help with illustrations.

REFERENCES

Air, G. M., Webster, R. G., Colman, P. M. & Laver, W. G. 1987 Distribution of sequence differences in influenza N9 neuraminidase of tern and whale viruses and crystallization of the whale neuraminidase complexed with antibodies. *Virology* **160**, 346–354.

Basler, C. F. (and 11 others) 2001 Sequence of the 1918 pandemic influenza virus nonstructural gene (NS) segment and characterization of recombinant viruses bearing the 1918 NS genes. *Proc. Natl Acad. Sci. USA* **98**, 2746–2751.

Beklemishev, A. B., Nazarova, L. M., Filimonov, N. G., Blinov, V. M., Grinev, A. A., Chuvakova, Z. K., Kim, E. V. & Mukazhanova, G. N. 1993 Synthesis, cloning, and determination of the primary structure of a full-length DNA copy of the gene for influenza A/Alma-Ata/1417/84 virus (H1N1-serovariant HSW1N1) hemagglutinin. *Mol. Gen. Mikrobiol. Virusol.* **1**, 24–27.

Bergmann, M., García-Sastre, A. & Palese, P. 1992 Transfection-mediated recombination of influenza A virus. *J. Virol.* **66**, 7576–7580.

Bikour, M. H., Frost, E. H., Deslandes, S., Talbot, B. & Elazhary, Y. 1995 Persistence of a 1930 swine influenza A (H1N1) virus in Quebec. *J. Gen. Virol.* **76**, 2539–2547.

Colman, P. M., Varghese, J. N. & Laver, W. G. 1983 Structure of the catalytic and antigenic sites in influenza virus neuraminidase. *Nature* **303**, 41–44.

Crosby, A. W. 1989 *America's forgotten pandemic: the influenza of 1918*. Cambridge University Press.

Duhaut, S. D. & Dimmock, N. J. 1998 Heterologous protection of mice from a lethal H1N1 influenza virus infection by H3N8 equine defective interfering virus: comparisons of defective RNA sequences isolated from the DI inoculum and mouse lung. *Virology* **248**, 241–253.

Fanning, T. G., Reid, A. H. & Taubenberger, J. K. 2000 Influenza A virus neuraminidase: regions of the protein potentially involved in virus–host interactions. *Virology* **276**, 417–423.

Fields, S. & Winter, G. 1982 Nucleotide sequences of influenza virus segments 1 and 3 reveal mosaic structure of a small viral RNA segment. *Cell* **28**, 303–313.

Gambaryan, A. S., Tuzikov, A. B., Piskarev, V. E., Yamnikova, S. S., Lvov, D. K., Robertson, J. S., Bovin, N. V. & Matrosovich, M. N. 1997 Specification of receptor-binding phenotypes of influenza virus isolates from different hosts using synthetic sialylglycopolymers: non-egg-adapted human H1 and H3 influenza A and influenza B viruses share a common high binding affinity for 6′-sialyl(N-acetyllactosamine). *Virology* **232**, 345–350.

Gibbs, A. J., Air, G. & Laver, G. 1982 Analysis of variation among haemagglutinin genes of influenza A viruses. In *Viral diseases in south-east Asia and the western Pacific* (ed. J. S. Mackenzie), pp. 546–549. Sydney: Academic Press.

Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. 2000 Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573–582.

Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. 2001*a* Recombination in the haemagglutinin gene of the 1918 'Spanish flu'. *Science* **293**, 1842–1845.

Gibbs, M., Armstrong, J. & Gibbs, A. 2001*b* SiScan version 2.0: Monte Carlo procedures for assessing signals in recombinant sequences. See http://life.anu.edu.au/molecular/software/siscan/siscan.htm.

Gorman, O. T., Bean, W. J., Kawaoka, Y., Donatelli, I., Guo, Y. J. & Webster, R. G. 1991 Evolution of influenza A virus nucleoprotein genes: implications for the origins of H1N1 human and classical swine viruses. *J. Virol.* **65**, 3704–3714.

Holmes, E. C., Worobey, M. & Rambaut, A. 1999 Phylogenetic evidence for recombination in dengue virus. *Mol. Biol. Evol.* **16**, 405–409.

Huang, Y. F., Shi, X. W. & Zhang, Y. P. 1999 Mitochondrial genetic variation in Chinese pigs and wild boars. *Biochem. Genet.* **37**, 335–343.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. 1998 Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* **23**, 403–405.

Khatchikian, D., Orlich, M. & Rott, R. 1989 Increased viral pathogenicity after insertion of a 28S ribosomal RNA sequence into the haemagglutinin gene of an influenza virus. *Nature* **340**, 156–157.

Kilbourne, E. 1977 Influenza pandemics in perspective. *J. Am. Med. Assoc.* **237**, 1225–1228.

Klenk, H. D. & Garten, W. 1994 Host cell proteases controlling virus pathogenicity. *Trends Microbiol.* **2**, 39–43.

Lai, M. M. C. 1992 RNA recombination in animal and plant viruses. *Microbiol. Rev.* **56**, 61–79.

Lederberg, J. 2001 H1N1-influenza as Lazarus: genomic resurrection from the tomb of an unknown. *Proc. Natl Acad. Sci. USA* **98**, 2115–2116.

McGuire, G., Wright, F. & Prentice, M. J. 1997 A graphical method for detecting recombination in phylogenetic data sets. *Mol. Biol. Evol.* **14**, 1125–1131.

Martindale, D. 2000 No mercy. *New Scient.* **168**, 28–32.

Maynard Smith, J. & Smith, N. H. 1998 Detecting recombination from gene trees. *Mol. Biol. Evol.* **15**, 590–599.

Morgan, D. J. & Dimmock, N. J. 1992 Defective interfering virus inhibits immunopathological effects of infectious virus in the mouse. *J. Virol.* **66**, 1188–1192.

Olsen, C. W., Carey, S., Hinshaw, L. & Karasin, A. L. 2000 Virologic and serologic surveillance for humans, swine and avian influenza virus infections among pigs in the north-central United States. *Arch. Virol.* **145**, 1399–1419.

Orlich, M., Gottwald, H. & Rott, R. 1994 Nonhomologous recombination between the hemagglutinin gene and the nucleoprotein gene of an influenza virus. *Virology* **204**, 462–465.

Pickrell, J. 2001 Killer flu with human–pig pedigree? *Science* **292**, 1041.

Reid, A. H., Fanning, T. G., Hultin, J. V. & Taubenberger, J. K. 1999 Origin and evolution of the 1918 'Spanish' influenza virus hemagglutinin gene. *Proc. Natl Acad. Sci. USA* **96**, 1651–1656.

Reid, A. H., Fanning, T. G., Janczewski, T. A. & Taubenberger, J. K. 2000 Characterization of the 1918 'Spanish' influenza virus neuraminidase gene. *Proc. Natl Acad. Sci. USA* **97**, 6785–6790.

Robertson, D. L., Sharp, P. M., McCutchan, F. E. & Hahn, B. H. 1995 Recombination in HIV-1. *Nature* **374**, 124–126.

Röhm, C., Zhou, N., Süss, J., Mackenzie, J. & Webster, R. G. 1996 Characterization of a novel influenza haemagglutinin, H15: criteria for the determination of influenza A subtype. *Virology* **217**, 508–516.

Saitou, N. & Nei, M. 1987 The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.

Sawyer, S. 1989 Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**, 526–538.

Signer, E. N., Dubrova, Y. E. & Jeffreys, A. J. 2000 Are DNA profiles breed-specific? A pilot study in pigs. *Anim. Genet.* **31**, 273–276.

Strimmer, K. & Von Haeseler, A. 1996 Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964–969.

Subbarao, K. (and 15 others) 1998 Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness. *Science* **279**, 393–396.

Tamura, K. & Nei, M. 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526.

Taubenberger, J. K., Reid, A. H., Krafft, A. E., Bijwaard, K. E. & Fanning, T. G. 1997 Initial genetic characterization of the 1918 'Spanish' influenza virus. *Science* **275**, 1793–1796.

Taubenberger, J. K., Reid, A. H. & Fanning, T. G. 2000 The 1918 influenza virus: a killer comes into view. *Virology* **274**, 241–245.

Varghese, J. N., Laver, W. G. & Colman, P. M. 1983 Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution. *Nature* **303**, 35–40.

Webster, R. G. 1999 1918 Spanish influenza: the secrets remain elusive. *Proc. Natl Acad. Sci. USA* **96**, 1164–1166.

Webster, R. G., Hinshaw, V. S. & Laver, W. G. 1982 Selection and analysis of antigenic variants of the neuraminidase of N2 influenza viruses with monoclonal antibodies. *Virology* **117**, 3–104.

Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. 1992 Evolution and ecology of influenza A viruses. *Microbiol. Rev.* **56**, 152–179.

Weiller, G. F. 1998 Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Mol. Biol. Evol.* **15**, 326–335.

Weiller, G. F. & Gibbs, A. J. 1993 DIPLOMO; distance plot monitor, v. 1.02. Available from ftp://life.anu.edu.au.

Weiller, G. F. & Gibbs, A. 1995 DIPLOMO: the tool for a new type of evolutionary analysis. *CABIOS* **11**, 535–540.

Wiley, D. C. & Skehel, J. J. 1987 The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *A. Rev. Biochem.* **56**, 365–394.

Worobey, M. & Holmes, E. C. 2001 Homologous recombination in GB virus C/hepatitis G virus. *Mol. Biol. Evol.* **18**, 254–261.

Worobey, M., Rambaut, A. & Holmes, E. C. 1999 Widespread intra-serotype recombination in natural populations of dengue virus. *Proc. Natl Acad. Sci. USA* **96**, 7352–7357.

Zhang, K., Hawken, M., Rana, F., Welte, F. J., Gartner, S., Goldsmith, M. A. & Power, C. 2001 Human immunodeficiency virus type 1 clade A and D neurotropism: molecular evolution, recombination, and coreceptor use. *Virology* **283**, 19–30.