Questioning the Evidence for Genetic Recombination in the 1918 "Spanish Flu" Virus

Influenza viral sequences have been obtained from preserved tissues of victims of the "Spanish flu" pandemic that killed over 20 million people from 1918 to 1919 (*1, 2*). Phylogenetic analysis of hemagglutinin (HA) gene sequences has indicated that the 1918 Spanish flu virus was more closely related to the human lineage than to the swine or avian influenza lineages of the H1N1 subtype (*2*). In a recent reanalysis of the 1918 HA gene, however, Gibbs *et al.* (*3*) proposed that recombination had occurred such that the majority of the globular domain (HA1) in the Spanish flu virus was acquired from the swine lineage, but its stalk region (HA2) was derived from the human lineage. Gibbs *et al.* also speculated that this intragenic recombination resulted in the increased virulence associated with the Spanish flu pandemic. We find no evidence for recombination in the 1918 HA gene, however. Rather, the apparent recombination described in (*3*) results from a difference in the rate of evolution between HA1 and HA2—a difference present only in human influenza A viruses.

In contrast to previous analyses (*1, 2*), Gibbs *et al.* (*3*) did not use the avian lineage to root their phylogenetic trees, and they suggested that the position of the bird lineage depends on the substitution model used. However, using avian influenza to distinguish the human and swine lineages results in phylogenetic tree topologies (Fig. 1) that are almost identical for both putative recombinant regions proposed by Gibbs *et al.* (*3*). In particular, the 1918 strain is placed on the human lineage not just for the putative human region but, crucially, for the supposed swine region as well (with 96% bootstrap support). Thus, maximum likelihood trees incorporating the avian lineage provide a robust, informative, and necessary test of the recombination hypothesis, whereas the midpoint rooting employed by Gibbs *et al.* is affected by differences in the rate of molecular evolution between the HA1 and HA2 gene regions. The phylogenies (Fig. 1) show the same relative depth of the swine influenza clades for HA1 versus HA2, illustrating a nearly constant rate of evolution across the entire HA in this lineage. In contrast, for the human lineage, the HA2 region evolves at a considerably lower rate than HA1. Pairwise comparisons among the human lineage strains confirm that rate difference, with uncorrected distances roughly twice as high for the HA1 region as for the HA2 region.

Plots presented by Gibbs *et al*. [figures 1A and B in (*3*)] give the strong impression that the 1918 strain is more similar to the swine

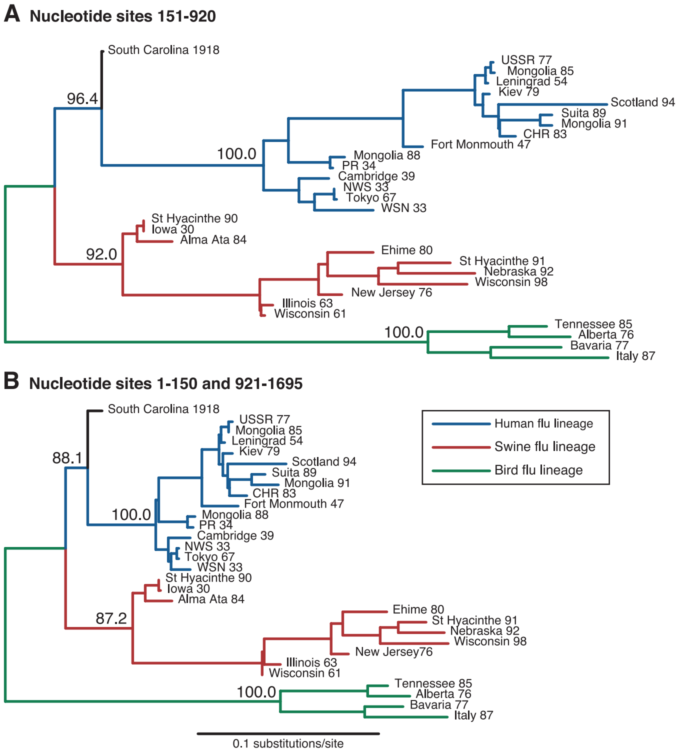

Fig. 1. Maximum-likelihood phylogenetic trees (*7*) for the recombinant regions in (**A**) HA1 and (**B**) HA2 identified by Gibbs *et al.* (*3*). The trees are drawn to the same scale. Compared with the swine lineage, which has apparently evolved at a similar rate in both regions, the human lineage has accumulated changes relatively rapidly in HA1 and relatively slowly in HA2.

Table 1. Variable sites in the recombinant regions in HA1 and HA2 identified by Gibbs *et al.* (*2*). *V*, number of variable sites; D_{human}, number of sites at which 1918 strain differs from human strain; D_{swine}, number of sites at which 1918 strain differs from swine strain.

	V	D_{human}	D_{human}/V (%)	D_{swine}	D_{swine}/V (%)
Globular domain region (sites 310-870)	142	94	66.2	61	43.0
Stalk domain region (sites 1070-1650)	95	45	47.4	64	67.4

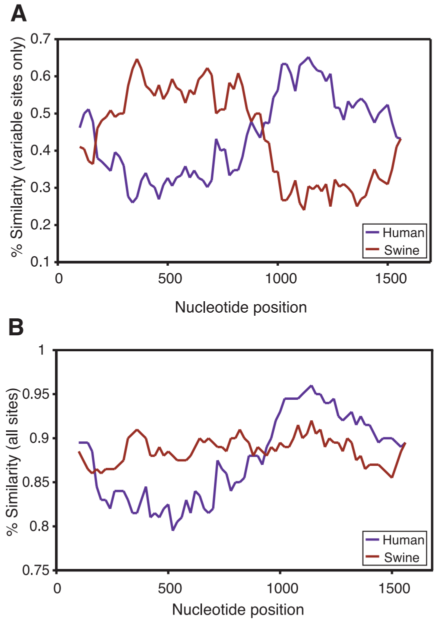

Fig. 2. Similarity plots of the complete HA gene of the 1918 strain with the human-lineage (Kiev 79) and swine-lineage (Wisconsin 61) reference sequences. *x* axis corresponds to nucleotide positions across the HA sequence alignment; *y* axis gives percent similarity. (**A**) Similarity at variable sites only, as plotted by Gibbs *et al*. (*3*). (**B**) Similarity at all sites.

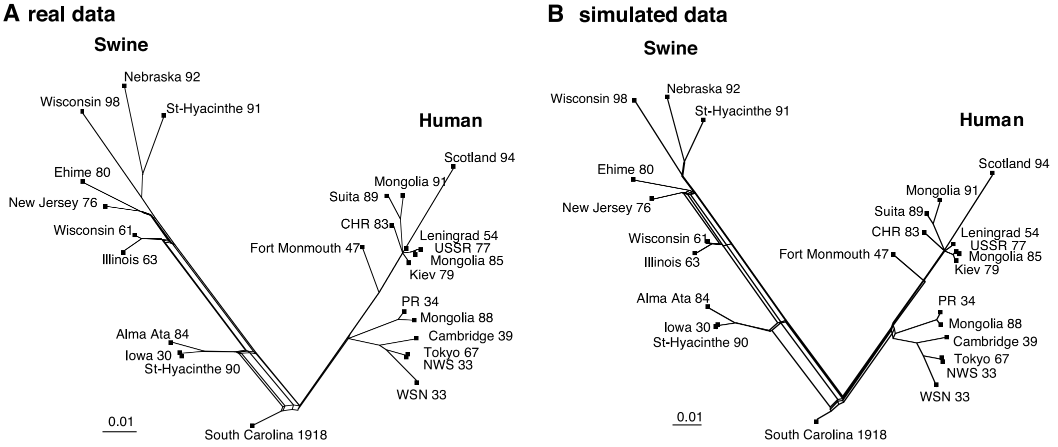

Fig. 3. Split-decomposition graphs (*4*) comparing (**A**) the human-lineage and swine-lineage complete HA gene sequences, and (**B**) sequences simulated without recombination.

lineage in the HA1 region than across the remainder of the HA gene. The 1918 sequence, however, actually exhibits about 11% uncorrected evolutionary distance to the swine reference strain across both regions (Fig. 2B). That fact is obscured because Gibbs *et al.* plotted the percent nucleotide identity at variable sites only, and the rate change in the human lineage will affect that proportion (Fig. 2A). There are considerably more variable sites in the HA1 region, where the human lineage evolves at a relatively high rate, than in the HA2 region, where the human lineage evolves at a relatively low rate (Table 1). Although there are about the same number of differences between the 1918 strain and the swine-lineage strain in each region, the similarity at variable sites is higher in the HA1 region than in the HA2 region. This effect will give the misleading impression of recombination, particularly when a relatively ancient strain that falls near the root of the two lineages is compared with reference strains from each lineage. The standard similarity plot, constructed using all sites (Fig. 2B), reveals that the 1918 HA sequence exhibits no greater affinity with the swine-lineage sequence in HA1 compared with HA2. The distance between the 1918 strain and the human-lineage reference strain, however, is about 17% in HA1 and 8% in HA2, a difference that is accounted for by the human-lineage rate bias evident in the phylogenies.

Finally, we simulated sequence data on a clonal tree, but incorporated the observed rate difference in the human lineage of HA1. Split-decomposition graphs (*4*) obtained from the simulated data were very similar to the graph obtained from the real data (Fig. 3), an indication that the rate difference across the different regions of the human-lineage HA gene—and not recombination—was the cause of the network pattern in the split-decomposition graph for the real data.

We conclude that there is no evidence that the HA gene of the 1918 Spanish flu virus had a recombinant origin. The finding of Gibbs *et al.* was the result of a localized difference in the rate of molecular evolution along the human-lineage HA gene. In view of the exceptional virulence of the 1918 epidemic, further investigations into the processes underlying the patterns of influenza sequence variability are required.

Michael Worobey
Andrew Rambaut
Oliver G. Pybus
David L. Robertson
Department of Zoology
University of Oxford
South Parks Road
Oxford, OX1 3PS, UK
E-mail: david.robertson@zoo.ox.ac.uk

References and Notes

1. J. K. Taubenberger, A. H. Reid, A. E. Krafft, K. E. Bijwaard, T. G. Fanning, *Science* **275**, 1793 (1997).
2. A. H. Reid, T. G. Fanning, J. V. Hultin, J. K. Taubenberger, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 1651 (1999).
3. M. J. Gibbs, J. S. Armstrong, A. J. Gibbs, *Science* **293**, 1842 (2001).
4. The split-decomposition analyses were performed using SplitsTree version 2.4 (*5*) with the default parameter settings. The simulated data set depicted (Fig. 3) was one of ten we produced using Seq-Gen (*6*). Sequences were evolved along the separate maximum likelihood trees reconstructed for each region, and the resulting alignments were then concatenated to produce data sets without recombination, but with the human-lineage rate differences in HA1 and HA2.
5. D. H. Huson, *Bioinformatics* **14**, 68 (1998).
6. A. Rambaut, N. C. Grassly, *Comput. Appl. Biosci.* **13**, 235 (1997).
7. Phylogenetic trees were reconstructed using PAUP* version 4 (*8*) under a general time-reversible (GTR) model with codon-position-specific rate heterogeneity. This model gave likelihoods greater by a factor of more than 50 than any of the nested models defined by ModelTest (*9*). A neighbor-joining tree was constructed and was used to estimate the parameters of the substitution process. These parameters were then used in a maximum likelihood heuristic search using tree bisection reconnection (TBR) followed by nearest neighbor interchange (NNI) branch-swapping. The parameters were then reestimated on this tree and the heuristic search repeated. The first and second heuristic searches gave the same tree for both regions. Bootstrapping was performed by producing 200 bootstrap replicates, sampling codons to preserve codon position rate heterogeneity. The bootstrap trees were constructed using a TBR heuristic search with the substitution model estimated as above. Selected relevant bootstrap values are displayed on the trees. The sequences used were the same as those listed in (*3*), and the regions included for each tree conformed to the recombinant fragments reported by Gibbs *et al.* (i.e., the partial HA1 region spanning nucleotides 151-920, and the remaining nucleotides, 1-150 and 921-1695, largely from the HA2, hereafter referred to as HA1 and HA2, respectively). Analyses on the smaller regions used for tree reconstruction in (*3*) (nucleotides 310-870 for HA1 and 1070-1650 for HA2) yielded similar results. To test the sensitivity of these results to the substitution model used, we did a maximum likelihood heuristic search using each of the 56 models available in PAUP* (*8, 9*) ranging from the simplest Jukes-Cantor to the general time-reversible with gamma distributed rate heterogeneity and a proportion of invariable sites. All resulting phylogenies placed the 1918 strain on the human lineage for the HA1 region [the region suggested to be on the swine lineage in (*3*)].
8. D. L. Swofford, PAUP* version 4.0b8 (Sinauer, Sunderland, MA, 2000).
9. D. Posada, K. A. Crandall, *Bioinformatics* **14**, 817 (1998).

22 December 2001; accepted 4 March 2002

Response: Worobey *et al.* point to evidence of constraining selection on the part of the HA gene that encodes the stem of the protein. This selection differentially affected the human but not the swine lineage of H1 influenzas. It had not been reported before and we had not recognized it, even though it was evident in our results. It affects a large part of the gene, which suggests that structural changes in the stem are selected against. Worobey *et al.* propose that the evidence of recombination we reported in the HA gene of the 1918 influenza (*1*) results from this differential selection.

This is an interesting idea, but they have not yet proven it. A phylogenetic analysis is not a "necessary test" for evidence of recombination (*2*). We doubt that the avian lineage may be used to root the tree of the mammalian lineages, because, like others (*3*), we found that the relative positions of the 1918 sequence and the avian lineage vary in trees, depending on what nucleotide sites or methods were used. We also found evidence that the HA sequences from birds and mammals are evolving quite differently, and we know of no way to prove which phylogenetic method or model is appropriate for this complex, incoherent data set (*2*). Current implementations of the maximum likelihood method used by Worobey *et al.* fit a single model to an entire dataset; thus, the positions of nodes (like the 1918 HA sequence) that link, or fall between, lineages that have evolved in different modes must be less than certain.

The degree of conservation and the strength of an evolutionary signal usually vary along a gene sequence, and it is a change in the relative strengths of several signals that is generally considered to indicate recombination (*4*). We used sister-scanning (*5*) to judge the relative support for the pairwise relatedness of four aligned sequences in a succession of windows sampled from along an alignment. Variations in the relationships between the swine and human lineage sequences were expected. By ignoring invariant sites, we eliminated one of the major sources of sequence variation that does not contribute anything to the relative signals. Relatedness can be expressed in various ways, but the values for each window are calculated quite independently of the others. Thus, our finding that, in a series of windows, the HA1 region of the 1918 HA gene sequence is significantly more closely related to the same region of swine lineage HA genes than to human lineage genes is independent of any features of the HA2 regions of the same sequences. There is no obvious reason to suspect that calculations showing the 1918 HA1 region to be more closely related to swine HA1 regions than human HA1 sequences were biased by differences in evolutionary rates.

The 1918 HA gene is most closely related to the HA gene of the oldest "classical swine" isolate (*3*), and we also found similar, although not identical, evidence of recombination in the HA gene sequence of that isolate. The proposal of Worobey *et al.* might explain the patterns found in the 1918 sequence, but it does not explain the patterns found in the HA gene of the classical swine isolate, which, for various reasons (*1*), we consider to be a member of the same recombinant lineage.

Mark J. Gibbs
John S. Armstrong
Adrian J. Gibbs
School of Botany and Zoology
The Australian National University
Canberra ACT 0200, Australia
E-mail: Mark.Gibbs@anu.edu.au

References

1. M. J. Gibbs, J. S. Armstrong, A. J. Gibbs, *Science* **293**, 1842 (2001).
2. ———, *Philos. Trans. R. Soc. London Ser. B* **356**, 1845 (2001).
3. A. H. Reid, T. G. Fanning, J. V. Hultin, J. K. Taubenberger, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 1651 (1999).
4. F. Gao *et al.*, *Nature* **397**, 436 (1999).
5. M. J. Gibbs, J. S. Armstrong, A. J. Gibbs, *Bioinformatics* **16**, 573 (2000).

24 January 2002; accepted 4 March 2002