JVI Accepts, published online ahead of print on 25 July 2007
J. Virol. doi:10.1128/JVI.00921-07

Persistent Host Markers in Pandemic and H5N1 Influenza Viruses

David B. Finkelstein[1], Suraj Mukatira[1], Perdeep K. Mehta[1], John C. Obenauer[1], Xiaoping Su[1], Robert G. Webster[2] and Clayton W. Naeve[1,3*]

[1]Hartwell Center for Bioinformatics and Biotechnology, St. Jude Children's Research Hospital, 332 North Lauderdale Street Memphis, TN 38105-2794, USA [2]Department of Infectious Diseases, St. Jude Children's Research Hospital, 332 North Lauderdale Street Memphis, TN 38105-2794, USA [3]Department of Pathology, University of Tennessee Health Science Center, Memphis, TN 38105, USA

Running Title: Persistent Host Markers in Influenza

Word Count: Abstract 181; Text 5,025

*** Reprint requests to:** C.W. Naeve, Ph.D.
Hartwell Center for Bioinformatics and Biotechnology
St. Jude Children's Research Hospital
332 North Lauderdale Street
Memphis, TN 38105-2794
Phone: 901-495-3689; Fax: 901-495-2945; E-mail: clayton.naeve@stjude.org

Key Words: Influenza, host, markers, H5N1, pandemic

ABSTRACT

Avian influenza viruses have adapted to human hosts causing pandemics in humans. The key host-specific amino acid mutations required for an avian influenza virus to function in humans are unknown. Through multiple sequence alignment and statistical testing of each aligned amino acid we identified markers that discriminate human influenza viruses from avian influenza viruses. We applied strict thresholds to select only markers which are highly preserved in human influenza isolates over time. We found that a subset of these persistent host markers exist in all human pandemic influenza sequences from 1918, 1957 and 1968, while others are acquired as the virus becomes a seasonal influenza. We also show that human H5N1 influenza viruses are significantly more likely to contain the amino acid predominant in human strains for a few persistent host markers when compared to avian H5N1 influenza viruses. This sporadic enrichment of amino acids present in human-hosted viruses may indicate that some H5N1 viruses have made modest adaptations to their new hosts in the recent past. The markers reported here should be useful in monitoring potential pandemic influenza viruses.

INTRODUCTION

The three best defined human influenza pandemics of the twentieth century may be derived in whole or part from avian influenza viruses (2, 47, 48), although the avian origin is disputed for the most deadly human pandemic known, the 1918 H1N1 "Spanish flu"(4, 15). This virus resulted in the deaths of millions of people worldwide (47). By comparison, avian H5N1 influenza viruses have killed 172 people since 1997 (http://www.who.int/). Despite containment efforts, H5N1 influenza infections of birds have spread across Asia to Europe, so that the potential for an H5N1 influenza

pandemic in humans still exists (21). Therefore, insight into the origin and adaptation of the 1918 H1N1 virus to humans may inform our understanding of the risks posed by H5N1 influenzas circulating in birds today.

Recently, large scale influenza genome projects have produced sufficient avian (34) and human (14) sequences to address fundamental questions. Given these resources, our aim was to determine precisely which amino acid changes best distinguish an avian influenza virus from a human influenza virus. After successfully identifying these amino acids, we used them to assess the significance of mutations in H5N1 influenza viruses isolated from humans. Furthermore, we defined a subset of these key amino acids which allowed us to track mutations in the H1N1 influenza lineage over time.

Although these human influenza viruses are independent isolates, they are not independent of lineage. The exact number of introductions is unknown, but these three influenza pandemics account for the overwhelming number of human influenzas and nearly all of the readily transmissible influenzas. As a result, a distinct amino acid from these three founding strains is more likely to have arisen by coincidence than the large sample size would suggest. Furthermore, the host and lineage parameters are so highly correlated that destratification methods based on PCA (39) or other methods (35) will erase the host effect. These methods have proven effective in other studies (39). However, our application of PCA based methods on the influenza sequence data failed to resolve host from lineage. The interpretation of host and lineage are therefore confounded. Specifically, we are precluded from determining whether host-differentiating amino acids are new adaptations or are due to the original lineages based on sequence data alone.

However, host markers that arose due to lineage may reasonably be of biological importance. As the successful colonization of human beings by influenza required the viruses to overcome selective pressure, even the original founding viruses of each lineage may reasonably be expected to contain important adaptations. Crucially, we can discern likely biologically significant host markers from those that are trivial by examining conservation. Since replication in influenza relies on low fidelity RNA polymerases (41), a high rate of random mutations is observed. Thus, given a large number of strains, we can estimate the expected frequency of amino acid substitution at a given position and compare that estimate to the observed frequency These estimates presume that the frequency of amino acid substitutions of the viruses do not vary substantially within a host. Variability in the amino acid substitution frequency between hosts is accounted for by our method (see Materials and Methods). Further, the influence of small violations of this assumption is moderated by the averaging effect of calculating the frequency of amino acid substitution across all influenzas within host.

Under this assumption, positions that are significantly more conserved than expected are likely to be important. In this context, we define conserved as having a low rate of amino acid substitution. Of course, we are not interested in those residues that are conserved in all influenzas generally. Rather, we are interested in residues that are conserved in a host-dependent manner. Thus, positions that are conserved only within humans are deemed biologically significant. Thus, even if these markers arose by chance or selection through the founding of the human influenza lineages, their high degree of persistence despite frequent mutation is evidence of biological relevance.

MATERIALS AND METHODS

Publicly available DNA and protein sequences were downloaded from the Influenza Virus Resource at NCBI (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) as of April 1, 2006. In addition, 847 newly sequenced avian influenza genes (Genbank accessions CY014548 - CY015177) were included. Sequences were retained if they began with methionine and were full length. Virus names were curated to conform to a fixed vocabulary. Ambiguities were manually verified or removed. Non-structural protein 2 (NS2), matrix protein 2 (M2) and polymerase basic protein 1 frame 2 product (PB1-F2) sequences were derived from NS, M and PB1 nucleotide sequences and translated for all downloaded sequences. NS1 sequences of lengths 217, 225, 230 and 237 amino acids were also included in the analysis. Qualified sequences were aligned using MUSCLE (12) and classified by serotype, country of origin, host and year. Bayesian analysis trees were generated to guide the manual editing of the HA and NA alignments (23). A total of 9,824 avian and 13,757 human influenza sequences were retained. Indonesian H5N1 human isolates (9) were also downloaded (http://flu.lanl.gov) (29) and included in the H5N1 population tests.

The data were reformatted for statistical testing such that each aligned position was in its own separate column and each row was from a single strain. Initial surveys revealed repeatedly sequenced examples of the same strain or of viruses from the same outbreak. To reduce this sampling bias, a representative set of genomes was selected. First, all sequences were classed by "outbreak", defined here as the set of all viruses with the same year, host, country and serotype. Next, the sequence which most closely matched the consensus of each outbreak was selected as the representative strain. This was repeated for each outbreak and for each gene. The resulting representative data set contained 6,561 protein sequences (Supplemental Information, data file).

We reasoned that H5N1, H7N7, and H9N2 isolates, recently introduced into humans, are not fully adapted, and thus may lack the persistent host markers we were seeking. Therefore these strains were excluded from defining host-specific residues, but were used to validate the results. Next, we identified the most frequent amino acid at each position in each gene of the avian and human isolates among the representative genes. Host specificity was tested at residues where the most frequent in avian influenza isolates differed from that in human influenza isolates. These positions were statistically tested (chi squared test using STATA 9.2/SE) for host specificity and the p-values were adjusted for multiple comparisons using the Bonferroni method. Next, the Euclidean distance between hosts at each position was calculated based on the frequency of each observed amino acid. A vector of amino acid frequencies from avian hosts was compared to a vector of amino acid frequencies from human hosts at each site. The Euclidean distance between these two vectors was then calculated for each site. This number was then divided by the square root of two, the maximum Euclidean distance possible, to create a proportion from zero to one that reflects the percentage of representative isolates that differ between host classes. A minimum proportional Euclidean distance of 0.95 was used as a threshold so that each host marker differentiated at least 95% of the representative isolates.

For each aligned position, the conservation of the most frequent amino acid found in human influenza isolates was calculated for both avian and human influenzas. Conservation was measured for all human H1N1, H2N2 and H3N2 influenza sequences, and for all avian sequences, not just the representative strains. Next, the conservation frequencies in human influenzas (x) were regressed against the conservation frequencies in avian influenzas (y) and standardized residuals were calculated. The regression line finds the overall difference in variability of influenzas between hosts. Extreme cases of

host dependent conservation have extreme standardized residuals of large absolute value. In this regression, large negative standardized residuals are those positions where conservation within human isolates is much larger than observed in birds. Based on z tables, the probability of a single marker having a standardized residual of less than -4 is .000032. Given 4,728 positions and 61 discoveries, a false discovery rate (7) of 1% was then calculated for positions which are below -4 standardized residuals. A position was deemed of interest if it passed the standardized residual threshold, had a 99% level of conservation in human influenza viruses and had a proportional Euclidean of 95% or greater. These positions were then designated as persistent host markers.

All full-length human influenza virus sequences of the correct serotype in the pandemic year were classed as pandemic viral sequences. For comparison purposes, an H5N1 consensus sequence was constructed by surveying all H5N1 viruses. The most frequent amino acid at each position in the pandemic strains and the H5N1 strains were then found. Pandemic markers are those host-differentiating sites where the most frequent amino acid is the same in all three pandemics.

RESULTS

Persistent host markers. We surveyed 9,824 avian influenza sequences including 847 novel avian genes sequenced by this laboratory and 13,757 human influenza sequences. We reduced sample bias by producing a representative data set of 6,561 sequences, minimized false positives by selecting markers with significant host-dependent conservation, and measured the persistence of changes over time (details in Methods and Supplemental Data). Using this approach we identified 32 amino acids, from 4,728 aligned positions, that distinguish avian and human influenza virus populations and that met all standards of host differentiation and host dependent conservation. Given a false

discovery rate of 1%, we expect that no more than 1 marker persists in human influenza viruses by chance. Given 611 sites that discriminate host to any degree, the median discriminator of host varies 205 times in 1000 human isolates. By comparison, each site we selected varies 10 times or less in 1000 human strains and simultaneously differentiates avian influenza viruses from human influenza viruses with 95% success.

These host markers are in five of the 11 proteins tested: RNA polymerase basic protein 2 (PB2), RNA polymerase acidic protein (PA), nucleoprotein (NP), matrix protein (M1), and the non-structural protein (NS1). The distribution of these residues among avian and human virus populations and the H1, H2 and H3 pandemic strains isolated during the first year of their respective pandemic is shown in Figure 1. The early 1918 (H1) pandemic isolates contain only 13 markers while the subsequent 1957 (H2) and 1968 (H3) pandemic strains contain all 32. This is likely due to the fact that these genes/proteins were derived from pre-existing human strains in the H2 and H3 pandemic viruses. A fourteenth marker, V100A in the PA protein, is shared by all pandemic strains but one and is thus not 100% conserved. H5N1 isolates, as a population, do not contain any of these markers although isolated cases do exist. By our stringent criteria, there are no persistent host markers in the surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA), or in the RNA polymerase basic protein 1 (PB1). This may seem surprising since avian derived HA, NA and PB1 genes have been identified in H2 and H3 pandemic isolates and one might expect to find host markers associated with these proteins, and particularly those residues in HA and NA selected to escape immune surveillance mechanisms (2). However, our methods are designed to identify strictly conserved residues that persist over time and will not capture seasonal changes or even changes between pandemic isolates (see Supplemental text).

Remarkably, 26 of the 32 markers (81%) are found in three of the four proteins that form the viral RNA replication complex (NP, PB2, and PA) (24, 44). Fourteen of these markers may be directly associated with the formation of the RNA replication complex as they fall in regions where NP, PB1 and PB2 are known to interact (37, 38) (Figure 2). Six markers in NP fall within known PB2 binding regions (38) and eight markers in PB2 are in regions of the molecule known to bind to either NP or PB1 (37). Two additional PB2 markers may influence RNA replication indirectly. The residue at 475 in the polymerase gene PB2 is predominately leucine (L) in avian isolates and methionine (M) human strains. This marker, L475M, is in a domain necessary for nuclear importation (30) and residue D567N is in the RNA cap binding region (20). The implication is that nuclear importation and formation of the RNA polymerase complex is influenced by the host environment. Less clear is the role PA plays in RNA replication and, consequently, the functional significance of the 10 markers in PA. One host marker, S225C, is located in a region involved in nuclear localization (32). The remaining 9 markers in PA are in regions of unknown or ambiguous functional importance.

The remaining six persistent host markers are found in M1 and NS1 proteins and are located in regions generally associated with binding to host cell proteins. All three host markers in M1 are in the C-terminal half of the molecule known to bind heat shock protein Hsc70 in host cells (49). Hsc70 has been shown to enhance viral replication through interaction with M1 (49). All three markers in the non-structural protein NS1 are located in regions of the molecule with known host cell binding functions. The N-terminal domain of NS1 binds to the 30 kDa subunit of the cleavage and polyadenylation specificity factor (CPSF) and to eukaryotic translation initiation factor 4 gamma 1 (eIF4G1) (5, 8) and contains the host marker I81M. In the course of identifying these markers in NS1 we identified a previously unreported SRC-homology 3 (SH3) motif. One of our host markers, P215T, is in this SH3 recognition motif where

the PPLPP motif is preserved in avian influenzas and is altered to PPLTP in human influenzas. The third persistent host marker in NS1 is at residue R227E in the PDZ (postsynaptic density, PSD-95; discs large, Dlg; zonula occludens-1, ZO-1) binding domain we previously identified at the C-terminus of the molecule and demonstrated to interact with numerous human PDZ domains *in vitro* (34). This region is also known to bind poly-A binding protein II, PABPII (11, 26). Given the NS1 protein's role in suppressing the host cell immune response through binding host proteins and host RNA, it may be that markers in this molecule point to key mutations needed to improve the immune suppression function of NS1 and enhance viral replication (25, 27).

Overall, persistent host markers are typically found in RNA replication complex proteins and are often located in known protein binding domains. These regions may directly influence the RNP replication complex or they may enhance replication through the interaction with host factors.

Persistent host markers in pandemic isolates. We next focused on the early isolates of pandemic influenza viruses to determine which markers they might have acquired. We found that 13 of our 32 host markers (Figure 1, arrows) are absolutely conserved (100%) in the influenza viruses that caused the 1918, 1957 and 1968 pandemics and are distributed among four viral genes; PB2, PA, NP and M1 (Table S1). Again the majority of these markers reside in RNA replication complex proteins. We should emphasize that it is unlikely that all 13 pandemic markers must be acquired to gain any single phenotypic trait of pandemic influenza viruses such as efficient replication, tissue tropism or transmissibility. Further, we cannot estimate how long it would take an avian virus such as H5N1 to acquire these traits.

DISCUSSION

Pandemic versus seasonal influenza. Although we cannot determine the rate at which avian isolates would acquire the 13 "pandemic" host markers, we can look at historical data to determine whether early pandemic isolates acquired additional markers over time. We can do this only for H1N1 isolates as they represent the introduction of all 8 influenza virus genes from an avian precursor, they have circulated in humans for 88 years, and many isolates have been sequenced. In contrast, subsequent pandemics involved the introduction of only HA, NA, and PB1 genes from an avian isolate into a pre-existing human strain, none of which carry host markers as defined by our criteria. If we plot the proportion of host markers in M1, NP, NS1, PA and PB2 proteins over time (Figure 3) we see that early H1N1 isolates, already containing 13 of the amino acids prevalent in human-hosted viruses, acquire the remaining 19 within 10-20 years depending on the protein. The stepped progression of PA and PB2 markers suggests that the H1N1 pandemic influenza adapted to human hosts in stages. In contrast, NS1 marker acquisition appears to be more abrupt. However, this is likely a sampling artefact as there are no H1N1 human influenza sequences available during the years 1919 to 1932. Further this abruptness may be due to the relatively few number of markers in NS1 and M1. Unlike these other genes, the host markers in proteins NP and M1 do not appear to be stably preserved, despite passing our 99% persistence criteria (see methods). The instability of these markers in these genes may be due to the reintroduction of H1N1 viruses from swine or birds or to seasonal variation in the human host.

The progressive changes seen in H1N1 human influenza isolates implies that these viruses gradually acquired mutations that confer the phenotypic traits of seasonal viruses and that these additional sites are not required for an influenza virus to cause a pandemic. Rather it is likely that these additional mutations are associated with the traits of seasonal influenza viruses such as low mortality. Over time, through successive rounds of transmission and selection, we would expect avian influenzas, like

H5N1, introduced into humans to acquire all 32 persistent host markers seen in seasonal influenza viruses.

Persistent host markers in H5N1 viruses. We examined H5N1 influenza sequences from avian hosts and compared them to H5N1 influenzas isolated from humans, focusing on the 32 persistent host markers. We included 7 H5N1 strains recently reported to transmit within a family in Indonesia (9). Although the predominant amino acid found in H5N1 isolates is consistent with avian influenzas at most marker locations, in a fraction of H5N1 isolates, the amino acid prevalent in human-hosted viruses has been acquired. We found four sites that are significantly enriched ($p < 0.0001$) in human H5N1 isolates (Table 1). Three of the four host markers that are enriched in H5N1 are also 100% conserved in human pandemic isolates. These three host markers are in PB2, one of which is the well known marker E627K. This mutation was seen in all seven of the putatively human-transmissible Indonesian H5N1 viruses (9) and in the 2003 H7N7 outbreak in the Netherlands (13). Four of the seven Indonesian strains also have the PB2 host marker K702R. This novel marker site is adjacent to a known high-pathogenicity site in PB2 at residue D701N (27). The enrichment of four host markers in H5N1 isolates suggests H5N1 influenza can adapt to human hosts. However, no single H5N1 virus sequenced contains more than two of these four sites.

The polymerase protein PB2 appears critical to adaptation of avian viruses to humans, based on this and other studies (9, 13). Significantly, we identify 10 PB2 host markers here (Table A1). These are all high quality discriminators of host (95% or greater) and all of these sites are preserved in 99% of human H1N1, H2N2 and H3N2 sequences over time. Not only does PB2, along with PA, have the most persistent host markers, it also has A199S, E627K, and K702R. These residues are the only host markers that are absolutely (100%) conserved in all pandemic influenza isolates we surveyed (Table A1)

and are also enriched in the population of human H5N1 isolates (Table 1). We suspect that acquisition of the amino acids that are prevalent in humans are required for the evolution of an avian influenza virus like H5N1 into a virus that is capable of causing a human pandemic. Here we must note that the sporadic and modest acquisition of markers in H5N1 human isolates and the stability of the H5N1 avian isolates indicate that currently circulating H5N1 viruses are no more adapted to human hosts today than they were in the past. What has changed is the geographic dispersion of the H5N1 virus and thus the size of the population at risk. Therefore the current risk of an H5N1 influenza pandemic in humans is due to an increased frequency of human exposure to the H5N1 virus from birds, rather than to a human adapted H5N1 virus.

Interestingly, two of these persistent host markers in PB2 occur in a unique set of four H5N1 human isolates from Indonesia. These Indonesian influenza isolates are distinguished from nearly all other human H5N1 isolates in that they may be acquired by human to human transmission rather than by avian to human transmission (9). Although the numbers are too small to allow a valid statistical test, these H5N1 isolates from a single Indonesian family appear more adapted to humans than the other H5N1 human isolates presumably acquired directly from birds. The residue A199S in PB2 is the only marker that is absolutely conserved in the seasonal human influenza isolates we surveyed (Table A1).

In summary, we have examined large collections of both avian and human influenza protein sequences and identified persistent host markers across the influenza proteome. By minimizing false positives and by focusing on those sites preserved over time in a host dependent manner, we have identified a set of 32 amino acids which are persistent host markers. These include both well known and novel sites, including a potential SH3 binding motif in NS1. By tracking the acquisition of these sites over time, we observed evidence of progressive adaptation of the avian H1N1 virus to human hosts. We show

306 that 13 of the 32 persistent host markers are 100% conserved amino acid changes in
307 pandemic viruses and suggest these are likely important in the evolution of pandemic
308 influenza. Further, we show that a small fraction of the population of H5N1 isolates
309 from humans have acquired four of these 32 markers, although no single H5N1 isolate
310 surveyed contains more than two markers and current H5N1 viruses are no more
311 adapted to humans today than they were in the past (Table 1).

312 **APPENDIX**

313 Details concerning bioinformatic and statistical methods are provided in this
314 supplement. The persistence percentages, and proportional Euclidean distances are
315 given in Table A1. A summary graphic of the proportional Euclidean distances by
316 genes is provided in Figure A1 and a frequency table of a key HA amino acid is shown
317 in Table A2. Further discussion of the statistical methods and results is provided
318 including a summary table of sample sizes in the representative set by gene. A brief
319 discussion of the protein interaction regions and the appropriate references are also
320 included here. The supplemental information includes an Excel file listing the
321 accession numbers, year, serotype and country of origin for all 6,561 representative
322 proteins.

323

324 **Multiple Sequence Alignments.** The protein sequences were aligned using the
325 MUSCLE (12) program. The MUSCLE program was chosen due to its performance,
326 flexibility, and the speed with which it aligns a large set of sequences. The protein
327 alignments were manually inspected and edited using BioEdit. Nucleotide sequences
328 were then aligned based on the protein alignments using the tranalign program in the
329 EMBOSS package. We found that protein-guided alignments of nucleotide sequences
330 produced better alignments than aligning the nucleotides directly. After generating

maximum likelihood trees, the sequences in each multiple alignment were re-ordered to match the ordering in the trees for easy visual comparison. The clade-guided re-alignment of nucleotide and protein sequences helped further improve the quality of alignments by manual checking and editing especially for HA and NA genes in the highly variable regions. Custom Perl scripts and additional EMBOSS tools were used to facilitate this process.

Potential for false negatives. One might expect, *a priori*, to find host markers in the surface glycoproteins HA and NA because of immune pressure and because of the receptor specificity of the HA receptor binding site (16, 40, 45, 50) or in the polymerase protein PB1 because of its association with HA/NA in the H2 and H3 pandemic strains. However, in this study, as a result of stringent criteria designed to eliminate false positives, authentic host adaptations may have been lost. As noted in the text, there are no host markers in the surface glycoproteins HA and NA or in the polymerase protein PB1. All amino acid markers from the genes HA, NA, and PB1 as well as the alternate transcripts NS2, M2 and PB1-F2 were either poor quality host discriminators (proportional Euclidean distance < 0.95) or were not preserved in human strains over time (persistence < 99%). Host specific residues in HA have been reported elsewhere (40), but HA residues do not differentiate more than 72% of viruses by host in this broad study (proportional Euclidean distance of 0.72). Two studies have also reported host-specific M2 sites (10, 28), however, these sites failed to pass thresholds used in this study. The best M2 site V86A (V is the predominant avian residue and A is the predominant human residue), did pass the Euclidean distance test, but failed the 99% persistence test. Thus, by our stringent criteria, this M2 site is a valid host-specific marker, but was excluded as a persistent host marker because it was not sufficiently preserved in human influenzas over time.

In addition, our methods test each residue separately, so that if host specific pressure can be relieved at any number of sites, then the pressure to conserve a given site is reduced as is the high degree of differentiation at that site. Direct evidence indicates that HA receptor specificity can be altered by mutations at any one of several sites(16, 40, 50). Furthermore our survey of all HA sequences indicates that the amino acids at key sites such as 226 in the HA receptor binding site are well preserved among avian influenza isolates, but are not well preserved among human influenza isolates (Table A2). We recognize that accurate alignments of HA and NA are hampered by high variability and despite the care taken in manual editing, false negative errors may occur due to alignment errors. While it might be possible to improve these alignments by adding structural data, this data only exists for portions of each protein and for only a few serotypes. Finally, while a lack of markers in HA and NA proteins are a concern, we note that there is also a lack of markers in PB1, which was trivial to align due to high conservation. Thus, it may be that residues in HA, NA, PB1 and PB1-F2 are simply less host-differentiating than are other genes, as we have observed.

Statistical Tests. All statistical tests in this paper are performed on categorical data. For each position we compared the frequency of amino acid categories across host using a two sided chi-squared test. This test assumed independence of the categories and is in common usage. For each position, the table size varied in accordance with the number of amino acid types. For the host test, we decided not to fix the table size at 2 by 20 to minimize table sparsity and to avoid false discoveries due to excessive degrees of freedom. We relied on the strictness of the Bonferroni correction and the application of absolute quality metrics to minimize false discoveries.

In total there were 4,728 aligned positions. Of these 611 positions passed the initial screening. Including the initial screens as informal hypothesis tests there were 4,728 tests. The Bonferroni threshold at the 0.05 alpha level was 1.06 e-05. There were 599

positions that passed Bonferroni criteria. As desrcibed in the methods section, applying the regression standardized residual threshold at -4 reduced this set to 61. This list was further reduced by the use of absolute metrics, proportional Euclidean distance and percent persistence. These metrics were chosen arbitrarily to guarantee minimum quality standards. Sample sizes were dependent on the gene of interest (Table A3).

The 32 remaining positions were then tested in H5N1 isolates for enrichment in human versus avian hosts. As this test generated two-by-two tables, Fisher's exact test was used. This two tailed test gives the most accurate p-value available under the independence assumptions and is computationally feasible for small tables. Again a Bonferroni threshold at the 0.05 alpha level was applied for these 32 tests. All results in Table 1 pass the Bonferroni threshold of 0.0015625. Sample size for the PB2 tests in Table 1 was 214 and the sample size for the PA test was 196.

Regions of Protein Interaction. Known regions of protein interaction in Figure 2 in the main text were derived from the literature. The M1 interaction regions are based on crystal structure (1) and viral assembly studies (3, 6, 18, 19, 31, 46). M1 interacts with other influenza viral proteins PA, NP, HA, PB1, and NS2 (18). M1 Also interacts with Hsc70 (36) . NS1 interacts with binds eukaryotic initiation factor 4GI at the N terminus (26), and has several protein binding domains (11, 25, 26), including a SH3 domain reported here for the first time (33). NP, PB1 (not pictured in Figure 2) and PB2 bind each other in RNA polymerase complex formation (37, 38). PA has casein kinase II sites (42), and binds to the RNA cap (17) and host proteins (22). PA also has a large proteolytic region (43) (from 1-247) that may interact with M1 (18); however, specific sites in PA are not known so this region was not included in the figure.

Accession Numbers. For Genbank or Genpept accession numbers of the genes included in the representative set are available on as a supplement.

ACKNOWLEDGEMENTS

This work was supported by the American Lebanese Syrian Associated Charities (ALSAC).

The authors gratefully acknowledge the technical comments and text editing of Caroline Obert, the editorial assistance of Geoff Neale, Yiping Fan, and Jinhua Wang and the statistical advice of Stan Pounds.

REFERENCES

1. **Akarsu, H., W. P. Burmeister, C. Petosa, I. Petit, C. W. Muller, R. W. Ruigrok, and F. Baudin.** 2003. Crystal structure of the M1 protein-binding domain of the influenza A virus nuclear export protein (NEP/NS2). Embo J **22:**4646-55.
2. **Alexander, D. J.** 2006. Avian influenza viruses and human health. Dev Biol (Basel) **124:**77-84.
3. **Ali, A., R. T. Avalos, E. Ponimaskin, and D. P. Nayak.** 2000. Influenza virus assembly: effect of influenza virus glycoproteins on the membrane association of M1 protein. J Virol **74:**8709-19.
4. **Antonovics, J., M. E. Hood, and C. H. Baker.** 2006. Molecular virology: was the 1918 flu avian in origin? Nature **440:**E9; discussion E9-10.

5. **Aragon, T., S. de la Luna, I. Novoa, L. Carrasco, J. Ortin, and A. Nieto.** 2000. Eukaryotic translation initiation factor 4GI is a cellular target for NS1 protein, a translational activator of influenza virus. Mol Cell Biol **20:**6259-68.

6. **Barman, S., A. Ali, E. K. Hui, L. Adhikary, and D. P. Nayak.** 2001. Transport of viral proteins to the apical membranes and interaction of matrix protein with glycoproteins in the assembly of influenza viruses. Virus Res **77:**61-9.

7. **Benjamini, Y., and Y. Hochberg.** 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological **57:**289-300.

8. **Burgui, I., T. Aragon, J. Ortin, and A. Nieto.** 2003. PABP1 and eIF4GI associate with influenza virus NS1 protein in viral mRNA translation initiation complexes. J Gen Virol **84:**3263-74.

9. **Butler, D.** 2006. Family tragedy spotlights flu mutations. Nature **442:**114-5.

10. **Chen, G. W., S. C. Chang, C. K. Mok, Y. L. Lo, Y. N. Kung, J. H. Huang, Y. H. Shih, J. Y. Wang, C. Chiang, C. J. Chen, and S. R. Shih.** 2006. Genomic signatures of human versus avian influenza A viruses. Emerg Infect Dis **12:**1353-60.

11. **Chen, Z., Y. Li, and R. M. Krug.** 1999. Influenza A virus NS1 protein targets poly(A)-binding protein II of the cellular 3'-end processing machinery. Embo J **18:**2273-83.

12. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792-7.

13. **Fouchier, R. A., P. M. Schneeberger, F. W. Rozendaal, J. M. Broekman, S. A. Kemink, V. Munster, T. Kuiken, G. F. Rimmelzwaan, M. Schutten, G. J. Van Doornum, G. Koch, A. Bosman, M. Koopmans, and A. D. Osterhaus.** 2004. Avian influenza A virus (H7N7) associated with human conjunctivitis and a fatal case of acute respiratory distress syndrome. Proc Natl Acad Sci U S A **101:**1356-61.

14. **Ghedin, E., N. A. Sengamalay, M. Shumway, J. Zaborsky, T. Feldblyum, V. Subbu, D. J. Spiro, J. Sitz, H. Koo, P. Bolotov, D. Dernovoy, T. Tatusova, Y. Bao, K. St George, J. Taylor, D. J. Lipman, C. M. Fraser, J. K. Taubenberger, and S. L. Salzberg.** 2005. Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. Nature **437:**1162-6.

15. **Gibbs, M. J., and A. J. Gibbs.** 2006. Molecular virology: was the 1918 pandemic caused by a bird flu? Nature **440:**E8; discussion E9-10.

16. **Glaser, L., J. Stevens, D. Zamarin, I. A. Wilson, A. Garcia-Sastre, T. M. Tumpey, C. F. Basler, J. K.**

Taubenberger, and P. Palese. 2005. A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity. J Virol **79:**11533-6.

17. **Hara, K., F. I. Schmidt, M. Crow, and G. G. Brownlee.** 2006. Amino acid residues in the N-terminal region of the PA subunit of influenza A virus RNA polymerase play a critical role in protein stability, endonuclease activity, cap binding, and virion RNA promoter binding. J Virol **80:**7789-98.

18. **Hara, K., M. Shiota, H. Kido, K. Watanabe, K. Nagata, and T. Toyoda.** 2003. Inhibition of the protease activity of influenza virus RNA polymerase PA subunit by viral matrix protein. Microbiol Immunol **47:**521-6.

19. **Harris, A., F. Forouhar, S. Qiu, B. Sha, and M. Luo.** 2001. The crystal structure of the influenza matrix protein M1 at neutral pH: M1-M1 protein interfaces can rotate in the oligomeric structures of M1. Virology **289:**34-44.

20. **Honda, A., K. Mizumoto, and A. Ishihama.** 1999. Two separate sequences of PB2 subunit constitute the RNA cap-binding site of influenza virus RNA polymerase. Genes Cells **4:**475-85.

21. **Horimoto, T., and Y. Kawaoka.** 2005. Influenza: lessons from past pandemics, warnings from current incidents. Nat Rev Microbiol **3:**591-600.

22. **Huarte, M., J. J. Sanz-Ezquerro, F. Roncal, J. Ortin, and A. Nieto.** 2001. PA subunit from influenza virus polymerase complex interacts with a cellular protein with homology to a family of transcriptional activators. J Virol **75:**8597-604.

23. **Huelsenbeck, J. P., and F. Ronquist.** 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17:**754-5.

24. **Kawaguchi, A., T. Naito, and K. Nagata.** 2005. Involvement of influenza virus PA subunit in assembly of functional RNA polymerase complexes. J Virol **79:**732-344.

25. **Krug, R. M., W. Yuan, D. L. Noah, and A. G. Latham.** 2003. Intracellular warfare between human influenza viruses and human cells: the roles of the viral NS1 protein. Virology **309:**181-9.

26. **Li, Y., Y. Yamakita, and R. M. Krug.** 1998. Regulation of a nuclear export signal by an adjacent inhibitory sequence: the effector domain of the influenza virus NS1 protein. Proc Natl Acad Sci U S A **95:**4864-9.

27. **Li, Z., H. Chen, P. Jiao, G. Deng, G. Tian, Y. Li, E. Hoffmann, R. G. Webster, Y. Matsuoka, and K. Yu.** 2005. Molecular basis of replication of duck H5N1 influenza viruses in a mammalian mouse model. J Virol **79:**12058-64.

28. **Liu, W., P. Zou, J. Ding, Y. Lu, and Y. H. Chen.** 2005. Sequence comparison between the extracellular domain of M2 protein human and avian influenza A virus provides new

information for bivalent influenza vaccine design. Microbes Infect **7:**171-7.

29. **Macken, C., H. Lu, J. Goodman, and L. Boykin.** 2001. The value of a database in surveillance and vaccine selection, p. 103-106. *In* A. D. Osterhaus, N. J. Cox, and A. W. Hampson (ed.), Options for the Control of Influenza IV. Elsevier Science, Amsterdam.

30. **Mukaigawa, J., and D. P. Nayak.** 1991. Two signals mediate nuclear localization of influenza virus (A/WSN/33) polymerase basic protein 2. J Virol **65:**245-53.

31. **Nayak, D. P., E. K. Hui, and S. Barman.** 2004. Assembly and budding of influenza virus. Virus Res **106:**147-65.

32. **Nieto, A., S. de la Luna, J. Barcena, A. Portela, and J. Ortin.** 1994. Complex structure of the nuclear translocation signal of influenza virus polymerase PA subunit. J Gen Virol **75 (Pt 1):**29-36.

33. **Obenauer, J. C., L. C. Cantley, and M. B. Yaffe.** 2003. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res **31:**3635-41.

34. **Obenauer, J. C., J. Denson, P. K. Mehta, X. Su, S. Mukatira, D. B. Finkelstein, X. Xu, J. Wang, J. Ma, Y. Fan, K. M. Rakestraw, R. G. Webster, E. Hoffmann, S. Krauss, J. Zheng, Z. Zhang, and C. W. Naeve.** 2006.

553 Large-scale sequence analysis of avian influenza isolates.
554 Science **311:**1576-80.
555 35. **Patterson, N., A. L. Price, and D. Reich.** 2006. Population
556 9Structure and Eigenanalysis. PLoS Genet **2:**e190.
557 36. **Perez-Gonzalez, A., A. Rodriguez, M. Huarte, I. J.**
558 **Salanueva, and A. Nieto.** 2006. hCLE/CGI-99, a human
559 protein that interacts with the influenza virus polymerase, is
560 a mRNA transcription modulator. J Mol Biol **362:**887-900.
561 37. **Poole, E., D. Elton, L. Medcalf, and P. Digard.** 2004.
562 Functional domains of the influenza A virus PB2 protein:
563 identification of NP- and PB1-binding sites. Virology
564 **321:**120-33.
565 38. **Portela, A., and P. Digard.** 2002. The influenza virus
566 nucleoprotein: a multifunctional RNA-binding protein
567 pivotal to virus replication. J Gen Virol **83:**723-34.
568 39. **Price, A. L., N. J. Patterson, R. M. Plenge, M. E.**
569 **Weinblatt, N. A. Shadick, and D. Reich.** 2006. Principal
570 components analysis corrects for stratification in genome-
571 wide association studies. Nat Genet **38:**904-9.
572 40. **Rogers, G. N., J. C. Paulson, R. S. Daniels, J. J. Skehel, I.**
573 **A. Wilson, and D. C. Wiley.** 1983. Single amino acid
574 substitutions in influenza haemagglutinin change receptor
575 binding specificity. Nature **304:**76-8.
576 41. **Sallie, R.** 2005. Replicative homeostasis II: influence of
577 polymerase fidelity on RNA virus quasispecies biology:

implications for immune recognition, viral autoimmunity and other "virus receptor" diseases. Virol J **2:**70.

42. **Sanz-Ezquerro, J. J., J. Fernandez Santaren, T. Sierra, T. Aragon, J. Ortega, J. Ortin, G. L. Smith, and A. Nieto.** 1998. The PA influenza virus polymerase subunit is a phosphorylated protein. J Gen Virol **79 (Pt 3):**471-8.

43. **Sanz-Ezquerro, J. J., T. Zurcher, S. de la Luna, J. Ortin, and A. Nieto.** 1996. The amino-terminal one-third of the influenza virus PA protein is responsible for the induction of proteolysis. J Virol **70:**1905-11.

44. **Sidorenko, Y., and U. Reichl.** 2004. Structured model of influenza virus replication in MDCK cells. Biotechnol Bioeng **88:**1-14.

45. **Stevens, J., O. Blixt, L. Glaser, J. K. Taubenberger, P. Palese, J. C. Paulson, and I. A. Wilson.** 2006. Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities. J Mol Biol **355:**1143-55.

46. **Takeuchi, H., A. Okada, and T. Miura.** 2003. Roles of the histidine and tryptophan side chains in the M2 proton channel from influenza A virus. FEBS Lett **552:**35-8.

47. **Taubenberger, J. K., A. H. Reid, R. M. Lourens, R. Wang, G. Jin, and T. G. Fanning.** 2005. Characterization of the 1918 influenza virus polymerase genes. Nature **437:**889-93.

48. **Tumpey, T. M., A. Garcia-Sastre, J. K. Taubenberger, P. Palese, D. E. Swayne, M. J. Pantin-Jackwood, S. Schultz-Cherry, A. Solorzano, N. Van Rooijen, J. M. Katz, and C. F. Basler.** 2005. Pathogenicity of influenza viruses with genes from the 1918 pandemic virus: functional roles of alveolar macrophages and neutrophils in limiting virus replication and mortality in mice. J Virol **79:**14933-44.

49. **Watanabe, K., T. Fuse, I. Asano, F. Tsukahara, Y. Maru, K. Nagata, K. Kitazato, and N. Kobayashi.** 2006. Identification of Hsc70 as an influenza virus matrix protein (M1) binding factor involved in the virus life cycle. FEBS Lett **580:**5785-90.

50. **Yamada, S., Y. Suzuki, T. Suzuki, M. Q. Le, C. A. Nidom, Y. Sakai-Tagawa, Y. Muramoto, M. Ito, M. Kiso, T. Horimoto, K. Shinya, T. Sawada, M. Kiso, T. Usui, T. Murata, Y. Lin, A. Hay, L. F. Haire, D. J. Stevens, R. J. Russell, S. J. Gamblin, J. J. Skehel, and Y. Kawaoka.** 2006. Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors. Nature **444:**378-82.

625 **TABLE 1.** Host markers are enriched in human H5N1 influenza viruses.

Gene	Mutation	Strains with a Human Adaptation Marker		p-value
		Avian H5N1 frequency %	Human H5N1 frequency %	
PB2	A199S	0/177 0%	7/37 19%	2.79E-06
	E627K	22/177 12%	20/37 54%	1.62E-07
	K702R	0/177 0%	6/37 16%	1.87E-05
PA	S409N	5/162 3%	11/34 32%	2.20E-06

626 P values are from Fisher's exact test.

627

628 **TABLE A1.** The 32 persistent host markers.

Gene	Position	Distance	Human	H1N1 1918	H2N2 1957	H3N2 1968	H5N1	Avian	Persistence
M1	115	0.967	I	V	I	I	V	V	99.27%
	121	0.962	A	A	A	A	T	T	99.92%
	137	0.958	A	T	A	A	T	T	99.10%
NP	16	0.953	D	D	D	D	G	G	99.41%
	61	0.973	L	I	L	L	I	I	99.32%
	283	0.981	P	P	P	P	L	L	99.24%
	305	0.96	K	R	K	K	R	R	99.07%
	313	0.973	Y	Y	Y	Y	F	F	99.32%
	357	0.964	K	K	K	K	Q	Q	99.32%
NS	81	0.958	M	I	M	M	deleted	I	99.32%
	215	0.955	T	P	T	T	P	P	.99.74%
	227	0.966	R	K	R	R	E	E	99.40%
PA	28	0.988	L	L	L	L	P	P	99.45%
	55	0.984	N	N	N	N	D	D	99.73%
	57	0.958	Q	R	Q	Q	R	R	99.27%
	100	0.955	A	A	A	A	V	V	99.54%
	225	0.969	C	S	C	C	S	S	99.36%
	268	0.951	I	L	I	I	L	L	99.09%
	337	0.978	S	A	S	S	A	A	99.91%
	404	0.967	S	A	S	S	A	A	99.27%
	409	0.959	N	S	N	N	S	S	99.54%
	552	0.999	S	S	S	S	T	T	99.91%
PB2	44	0.966	S	A	S	S	A	A	99.11%
	64	0.954	T	M	T	T	I	M	99.82%
	199	0.997	S	S	S	S	A	A	100.00%
	271	0.958	A	T	A	A	T	T	99.38%
	475	0.994	M	M	M	M	L	L	99.91%
	567	0.977	N	N	N	N	D	D	99.29%
	588	0.971	I	A	I	I	A	A	99.38%
	627	0.977	K	K	K	K	E	E	99.82%
	674	0.969	T	A	T	T	A	A	99.47%
	702	0.955	R	R	R	R	K	K	99.38%

629 Note each column has the most frequently occurring amino acid by class. Position is the
630 location in the protein sequence. Distance refers to the proportional Euclidean distance
631 of amino acid frequency between human and avian hosted viruses.

632 **TABLE A2.** The frequency of residues at position 226 varies by host in influenza A

633 haemagglutinin.

Amino Acid	Avian virus	Human virus	Row total
I	2	220	222
L	49	125	174
M	1	0	1
P	1	0	1
Q	1007	298	1305
R	0	1	1
V	2	626	628
ambiguity	0	2	2
deletion	1	0	1
Column total	1063	1272	2335

634

635 **TABLE A3.** Sample sizes by gene for statistical testing.

636

Gene	Sample size
HA	606
M1	697
M2	690
NA	683
NP	573
NS1	681
NS2	699
PA	481
PB1	490
PB1F2	481
PB2	480

FIGURE LEGENDS

Fig 1. Host differentiating sites are compared to pandemic strains. Each of the 32 host-differentiating sites are displayed and color-coded by host. Avian is in blue, human in yellow. The intensity of each position is determined by the proportional Euclidean distance between hosts. Positions where the consensus residue of each pandemic strain agrees with the most frequent human amino acid are boxed. The 13 positions where all pandemic isolates surveyed absolutely agree with the most frequent human amino acid are denoted by a black arrow. Wherever the most frequently observed amino acid is neither the avian nor human consensus residue, it appears in gray. Position numbers for the markers in each protein are given in the final row.

Fig. 2. Persistent host markers occur in known protein binding domains. Blue squares denote regions where the named protein is known to bind to a specific protein (Supplemental Information, text) or the novel SH3 domain. Red lines denote host markers found in this study.

Fig. 3. The preservation of host markers increases over time in human H1N1 viruses. The panels above plot the proportion of host markers acquired over time by H1N1 influenza isolated from human hosts. All 32 markers are 99% persistent. Position numbered in red are referred to in the text.

Fig. 1

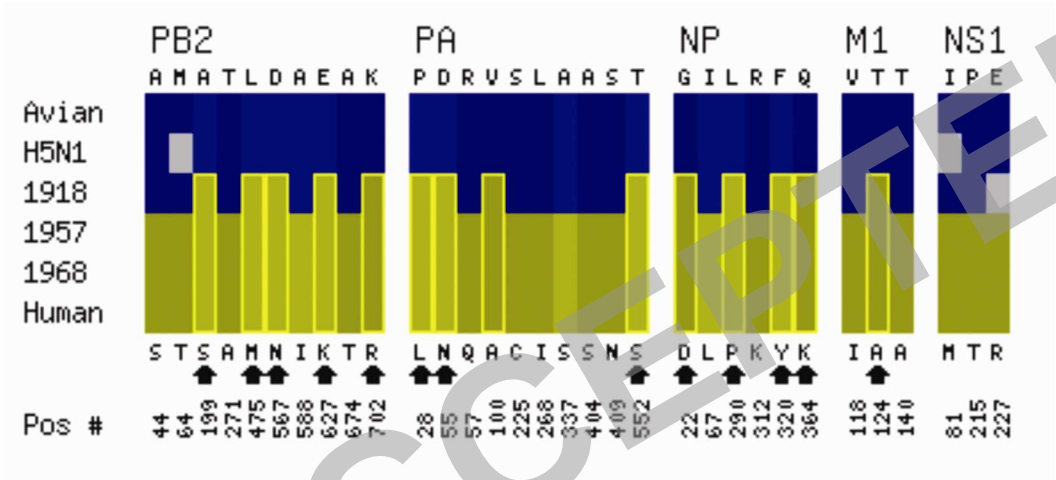

Fig. 2

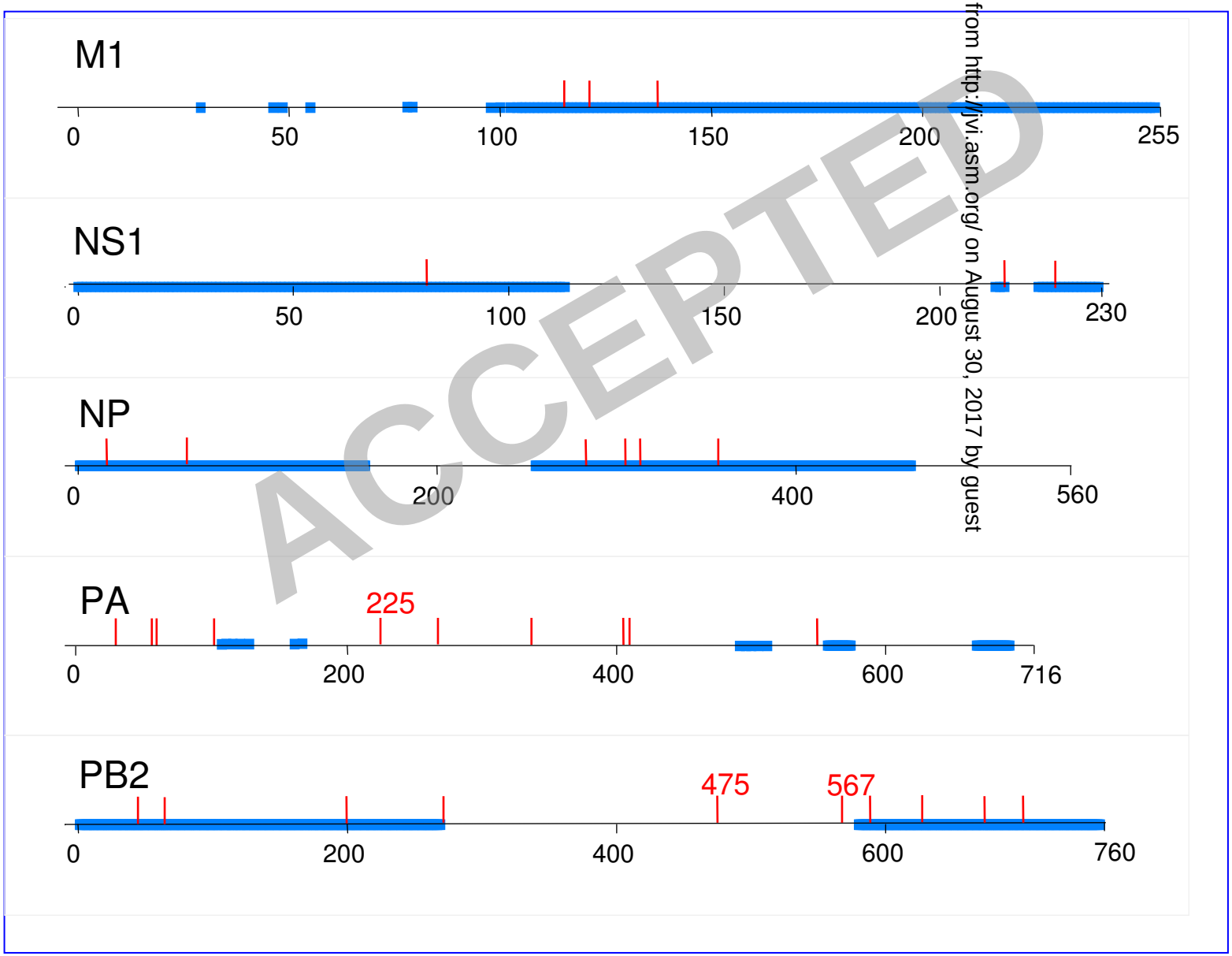

Fig. 3