



OPEN

Survival analysis of infected mice reveals pathogenic variations in the genome of avian H1N1 viruses

SUBJECT AREAS:

COMPUTATIONAL
BIOLOGY AND
BIOINFORMATICS

ECOLOGY

Zeynep A. Koçer¹, Yiping Fan², Robert Huether^{2*}, John Obenauer^{2†}, Richard J. Webby¹, Jinghui Zhang², Robert G. Webster¹ & Gang Wu²

Received

3 October 2014

Accepted

24 November 2014

Published

12 December 2014

Correspondence and requests for materials should be addressed to R.G.W. (robert.webster@stjude.org) or G.W. (gang.wu@stjude.org)

* Current address: Ambry Genetics, 15 Argonaut, Aliso Viejo, California 92656, United States.

† Current address: Digital Genomics LLC 2151 Dickens Place Drive, Southaven, Mississippi, 38672, United States.

¹Department of Infectious Diseases, Division of Virology, St. Jude Children's Research Hospital, Memphis, Tennessee, 38105, United States, ²Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, Tennessee, 38105, United States.

Most influenza pandemics have been caused by H1N1 viruses of purely or partially avian origin. Here, using Cox proportional hazard model, we attempt to identify the genetic variations in the whole genome of wild-type North American avian H1N1 influenza A viruses that are associated with their virulence in mice by residue variations, host origins of virus (Anseriformes-ducks or Charadriiformes-shorebirds), and host-residue interactions. In addition, through structural modeling, we predicted that several polymorphic sites associated with pathogenicity were located in structurally important sites, especially in the polymerase complex and NS genes. Our study introduces a new approach to identify pathogenic variations in wild-type viruses circulating in the natural reservoirs and ultimately to understand their infectious risks to humans as part of risk assessment efforts towards the emergence of future pandemic strains.

Influenza A viruses (IAVs), which belong to the *Orthomyxoviridae* family, typically replicate in the intestinal tract of wild aquatic birds causing no overt disease and are transmitted through open-water sources via the fecal-oral route¹. However, in mammals IAVs cause mainly respiratory infections. The genome of IAVs is a single-stranded, negative-sense RNA that consists of eight segments coding up to 12 proteins (PB2, PB1, PB1-F2, PA, PA-X, HA, NP, NA, M1, M2, NS1, and NEP)²⁻⁵. To date, there are 18 hemagglutinin (HA) and 11 neuraminidase (NA) subtypes of IAVs^{1,6-9}.

Although almost all combinations of HA and NA subtypes are found in wild aquatic birds, only H1, H2, and H3 subtypes have caused epidemics and pandemics in the human population. H1N1 IAVs that descended from avian sources probably moved into the human and swine populations prior to the 1918 pandemic and have established in those hosts since then^{10,11}. During reported human history, three pandemics have been caused by H1N1 viruses in 1918, 1977, and 2009. In particular, the pandemics in 1918 and 2009 resulted from the direct adaptation of the virus to mammals or the acquisition of several segments from avian H1N1 viruses prior to mammalian adaptation. Recently, we showed that the naturally circulating avian H1N1 viruses of North America could lead to high mortality in mice¹². Therefore, it is important to understand the genomic characteristics of these wild-type avian H1N1 viruses that facilitate interspecies transmission and pathogenesis in mammals without prior adaptation in another mammalian host.

To date, several molecular markers or genomic changes have been identified in the genomes of IAVs in terms of mammalian adaptation and increased virulence upon transmission to a new host. For instance, the E627K mutation in the PB2 gene¹³ was detected in recently emerged H7N9 human influenza A isolates¹⁴. However, the natural existence of such markers in the circulating IAV population of wild bird origins is not well known due to lack of surveillance studies and the corresponding genomic data. With the increasing number of surveillance studies across the world, it is important to investigate the unique genomic variations in some of the avian IAVs that make them virulent and transmissible in mammals; so we can be prepared for future outbreaks and pandemics originated from wild bird population.

In light of these concerns, we sequenced the whole genome of 31 wild-type North American avian H1N1 IAVs that were used in our previous study¹². Empowered by the whole genome information, in this study, we applied Cox proportional hazard model to the survival data that we previously obtained from mice infected by these viruses (n = 5 mice/virus) to evaluate the effects of deduced amino acid (aa) residue variation at each position,



Table 1 | Distribution of polymorphic sites in the proteome of the tested North American avian H1N1 IAVs

Origin	PB2	PB1	PB1-F2	PA	PA-X	HA	NP	NA	M1	M2	NS1	NEP	Genome
Number of polymorphic sites in the proteome	37	34	48	45	12	49	22	44	2	9	80	22	404
Polymorphic sites associated with pathogenicity by residue effect	8	3	6	5	-	2	-	2	-	1	60	18	105
Polymorphic sites associated with pathogenicity by host effect	4	5	7	1	-	2	2	-	-	-	1	1	23
Polymorphic sites associated with pathogenicity by host-residue interaction	-	1	4	3	4	-	-	-	-	-	1	1	14

The polymorphic sites associated with pathogenicity were identified by Cox proportional hazard model. Adjusted p value (FDR) ≤ 0.01 was deemed significant. These positions were based on the sequences used in this study.

host origin for viruses (Anseriformes or Charadriiformes), the interaction between residue and virus host. In many medical studies, time to death is the main event of interest. It is common that at the end of follow-up time some of the individuals have not had the event of interest, thus their true time to event is unknown. Furthermore, survival data are rarely normally distributed. These features of the data required a special method called survival analysis. In study designed to determine multiple factors that influence the survival, the most commonly used regression model is the Cox proportional hazard model. The Cox proportional hazard model has been used in Genome-wide association study^{15,16} to identify the genetic variants that are associated with survival of patients. This is a semi-parametric model as we model the covariate effects with a parametric model, but we model the baseline hazard rate (common to all samples) non-parametrically. After excluding the least pathogenic virus in DBA/2J mice due to only 20% overall mortality, this statistical model analyzes survival data to identify differences in survival due to independent variable such as treatment (i.e. virus infection) while adjusting other influential covariates (i.e. host origin), providing an estimate of the hazard ratio and its confidence interval. We then assessed the potential contribution of a subset of significant polymorphic sites to the virulence of the virus by structurally modeling and predicting their functional impact.

Results

Polymorphic sites observed in 31 wild-type North American avian H1N1 IAVs. We performed whole genome sequencing on 31 wild-type North American avian H1N1 viruses and deduced their protein sequences for each segment that were then aligned in order to compare their variations in the deduced proteome.

In total, we observed 404 (8.7%) variable sites in the full proteome of avian H1N1 IAVs (Table 1). The variable sites across the whole genome for all 31 viruses were shown in a proteotype plot (Fig. 1). NS1 contributed the most (80) to these variable sites, followed by HA (49), PB1-F2 (48), PA (45) and NA (44) (Table 1). Among all proteins, PB1-F2, NS1, PA-X, NEP have over 10% aa showing variations (53.3%, 34.8%, 19.7%, and 18.2%, respectively, Table 1). The individual plots showing polymorphic sites and the residue variations in each coding region are given in Supplementary Figs. S1–S11.

Using previously determined survival data in DBA/2J mice upon infection with wild-type North American avian H1N1 IAVs¹², we then investigated the effects of the residue variations, host origin of viruses, and host-residue interaction in relation to their pathogenicity in mice.

Effect of residue variation alone. After adjusting the host effect, the survival analysis with Cox proportional hazard model identified 108

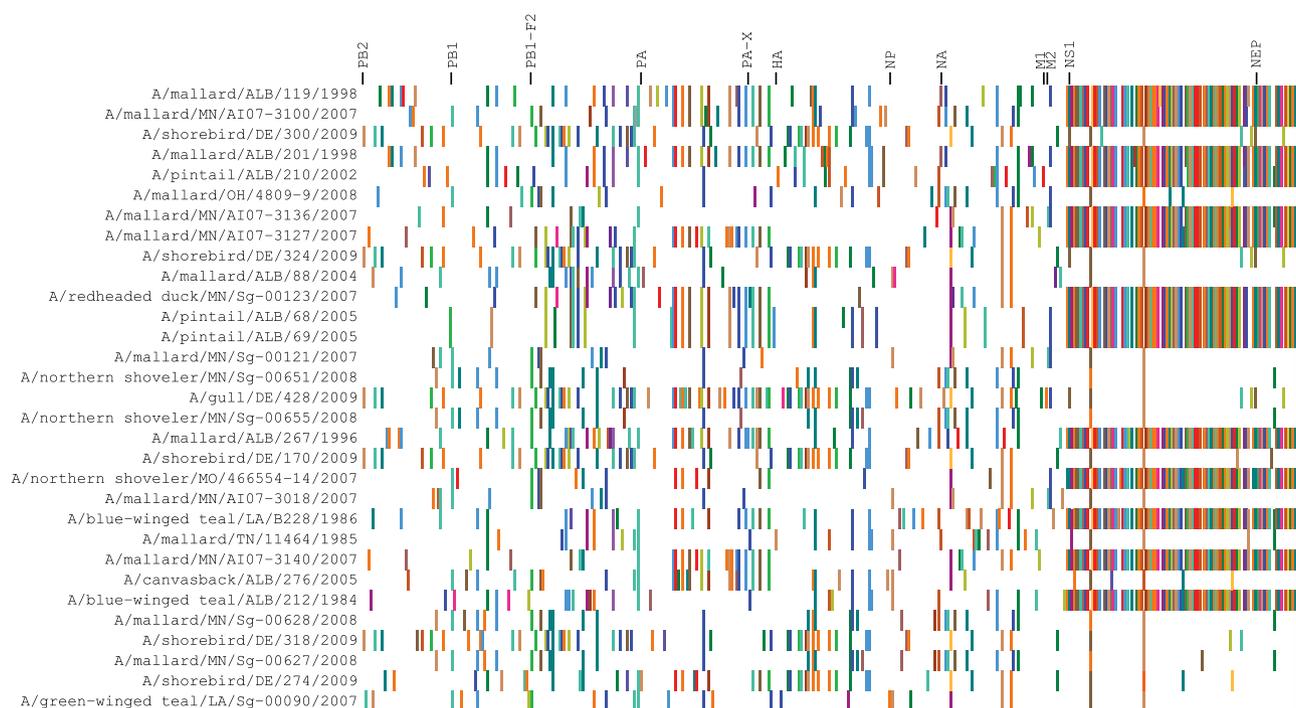


Figure 1 | Distribution of polymorphic residues in the proteome of North American avian H1N1 IAVs. The proteome was concatenated using the polymorphic sites in each protein in the following order: PB2, PB1, PB1-F2, PA, PA-X, HA, NP, NA, M1, M2, NS1, and NEP. Similar amino acids were grouped by color: aliphatic amino acids, shades of brown and orange; positively charged residues, shades of blue; negatively charged ones, shades of red; polar uncharged ones, shades of green; aromatic ones, shades of purple; and glycine and cysteine, shades of yellow. These positions were based on the sequences used in this study.



polymorphic residues at 105 sites (26% of the total polymorphic sites) that were statistically associated with the pathogenicity of the viruses in the DBA/2J mice by residue effect (adjusted P (FDR) \leq 0.01; Supplementary Tables S1, S2). The polymorphic sites with residue effect were contributed by NS1 (60), NEP (18), PB2 (8), PB1-F2 (6), PA (5), PB1 (3), HA (2), NA (2), and M2 (1) (Table 1, and Supplementary Table S2). Three of 105 polymorphic sites had more than one substitution at each site (R75H or H75L in PB1-F2, S7L or T7L in NS1, and S7L or T7L in NEP) (Supplementary Table S2). None of the variable sites in PA-X, NP, or M1 proteins were associated with the pathogenicity of the virus strains in terms of amino acid substitutions.

Notably, among these 105 sites, 78 (74.3%) were from NS1 and NEP, which were coded as spliced products by NS segment. The proteotype of NS1 and NEP indicates that there exist two major alleles in these 31 wild-type H1N1 viruses (Fig. 1, Supplementary Figs. S10, S11). In other words, the residue effect in these two proteins was mainly contributed by the underlying alleles rather than the individual variable sites. Therefore, we adjusted the effect of these non-independent polymorphic sites. After removing the 73 sites from NS1 and NEP that do not vary from the underlying allele of NS, the remaining 5 significant sites in these two proteins are 7, 27, 227 in NS1, and 7 and 70 in NEP. Due to the overlapping NS1 and NEP ORFs, the nucleotide substitutions in NS segment leading to residue variations in NS1 and NEP are: c.C20T for S7L in NS1 and NEP, c.A19T, c.C20T and c.C21A for T7L in NS1 and NEP, c.A81G for I27M in NS1, c.A680G for E227G in NS1 and S70G in NEP. With this justification, the total number of significant polymorphic sites associated with pathogenicity of these viruses in mice is reduced to 32 (7.9% of the 404 polymorphic sites).

Among the remaining 32 sites, the most significant residue variations associated with increased pathogenicity were K43R and D416N in NA, K53R and R75H in PB1-F2, I292T and E358V, S203N and I550L in HA, S7L in NS1 and NEP, S70G in NEP, E227G in NS1 and L187I in PA (Fig. 2; adjusted P (FDR) \leq 4.32E-05). For instance, the viruses with the residue N416 in their NA cause 100% mortality by 6 days post infection (dpi); however, the ones with the residue D416 in their NA showed delayed mortality (Fig. 2a).

Effect of host origin of virus. To understand whether the host origin of an avian virus (i.e. Anseriformes or Charadriiformes) is associated with its pathogenicity in mice, we investigated the host effect after adjusting the residue effect. Thus, the host effect was determined between the IAVs of Anseriformes and Charadriiformes origin carrying only the same residue at those sites. Twenty-three polymorphic sites exhibited an effect of the host origin of virus on the pathogenicity (adjusted P (FDR) \leq 9.96E-03; Supplementary Table S3). The polymorphic sites affected by host origin of viruses were in PB1-F2 (7), PB1 (5), PB2 (4), NP (2), HA (2), PA (1), NS1 (1), and NEP (1) (Table 1, and Supplementary Table S3). For instance, H75 in PB1-F2 and S70 in NEP were present in H1N1 viruses of Anseriformes and Charadriiformes origin but showed increased pathogenicity in mice infected with H1N1 viruses of Anseriformes origin (Fig. 3a, b).

Effect of host-residue interactions. We then examined the host-residue interaction and identified 14 such polymorphic sites (adjusted P (FDR) \leq 0.01) based on the residue effect on survival in different hosts. These sites were distributed mainly in polymerase genes (Table 1 and Supplementary Table S4). For example, the substitutions M46T in PB1-F2 in H1N1 IAVs of Charadriiformes origin was associated with increased mortality in the infected mice. However, the same substitution M46T in the viruses of Anseriformes origin showed decreased mortality in infected mice (Fig. 3c). On the other hand, the presence of I89V in NEP substitutions in viruses of Anseriformes origin was associated with increased mortality in infected mice, but decreased mortality in mice infected with the

viruses of Charadriiformes origin harboring the same substitutions (Fig. 3d).

Polymorphic sites associated with pathogenicity in individual proteins. The pathogenesis of IAVs is a complex polygenic trait that is an outcome of various molecular variations in the overall gene constellation. More importantly, some of those variations cause alteration in protein structure and function that might contribute to the efficacy of viral entry, replication and transmission in different host systems. Therefore, we examined the polymorphic sites in individual proteins, mainly the ones statistically associated with pathogenicity, in terms of their structural and functional importance. The presence of known adaptation, virulence or drug resistance markers and structural importance of all other observed genetic variants, where possible, are given in *Supplementary Information*.

Polymerase proteins. Being a component of the viral polymerase complex, the PB2 protein contributes to the efficiency of viral replication in different hosts. Of the 37 polymorphic sites in the PB2 protein (Supplementary Fig. S1), eight (I67V, A152S, A199T, V255I, I292T, E358V, R508Q, and V649I) were associated with increased pathogenicity by residue effect (adjusted P (FDR) \leq 4.77E-03; Supplementary Table S2). Four sites (67, 152, 199, and 508) were associated with pathogenicity of avian H1N1 IAVs by host effect (adjusted P (FDR) = 9.50E-03; Supplementary Table S3). Five of those sites with residue effect (I67V, A152S, A199T, R508Q, and V649I) were observed to co-occur in five avian H1N1 IAVs of Charadriiformes origin (A/shorebird/DE/300/2009, A/shorebird/DE/324/2009, A/gull/DE/428/2009, A/shorebird/DE/170/2009, and A/shorebird/DE/318/2009). Two variants associated with increased pathogenicity, E358V and V649I [adjusted P (FDR) are 5.63E-09 (Fig. 2e) and 4.77E-03, respectively] are located in the cap-binding and the host-specific domain of PB2 (Table 2). These variants were observed in some of the most pathogenic viruses (pathogenicity index = 4, ref. 12) and in five H1N1 IAVs of Charadriiformes origin mentioned above (Supplementary Fig. S1). The variant A199T, also observed only in viruses of Charadriiformes origin, was identified in association with pathogenicity by residue effect (adjusted P (FDR) = 4.77E-03). Interestingly, we also detected host effect at the same site when the viruses carrying A199 were compared (adjusted P (FDR) = 9.50E-03). This site has been previously identified as one of the genomic markers differentiating human and avian IAVs¹⁷.

Of the 34 polymorphic sites in the PB1 protein (Supplementary Fig. S2), we identified three substitutions (K215R, L298I, and I667V) associated with increased pathogenicity in mice by residue effect (adjusted P (FDR) \leq 4.77E-03; Supplementary Table S2), two of which (L298I and I667V) co-occurred in only five avian H1N1 IAVs of Charadriiformes origin mentioned above. Additionally, we observed five polymorphic sites in PB1 (215, 298, 372, 642, and 667) that were associated with pathogenicity by host effect (adjusted P (FDR) = 9.50E-03; Supplementary Table S3). Based on host-residue interactions, the S375N substitution was associated with increased pathogenicity in avian H1N1 IAVs of Anseriformes origin, and the N375S was more pathogenic in those of Charadriiformes origin (adjusted P (FDR) = 8.66E-03; Supplementary Table S4).

PB1-F2 is a small polypeptide expressed from the second ORF of the PB1 gene, and contributes to the viral fitness by inducing apoptosis in the infected host cell². We observed that 30 North American avian H1N1 viruses possess a full-length (90-aa) PB1-F2 ORF; but A/mallard/MN/AI07-3100/2007 possesses a slightly shorter (87-aa) PB1-F2 ORF. Of the 48 polymorphic sites in the PB1-F2 protein (Supplementary Fig. S3), seven residues (H15R, N23S, T27I, K53R, L58S, R75H, H75L) were associated with pathogenicity in mice by residue effect (adjusted P (FDR) \leq 4.77E-03; Supplementary Table S2). Four of the substitutions (N23S, T27I, L58S, and H75L) co-occurred in five avian H1N1 IAVs of Charadriiformes origin men-

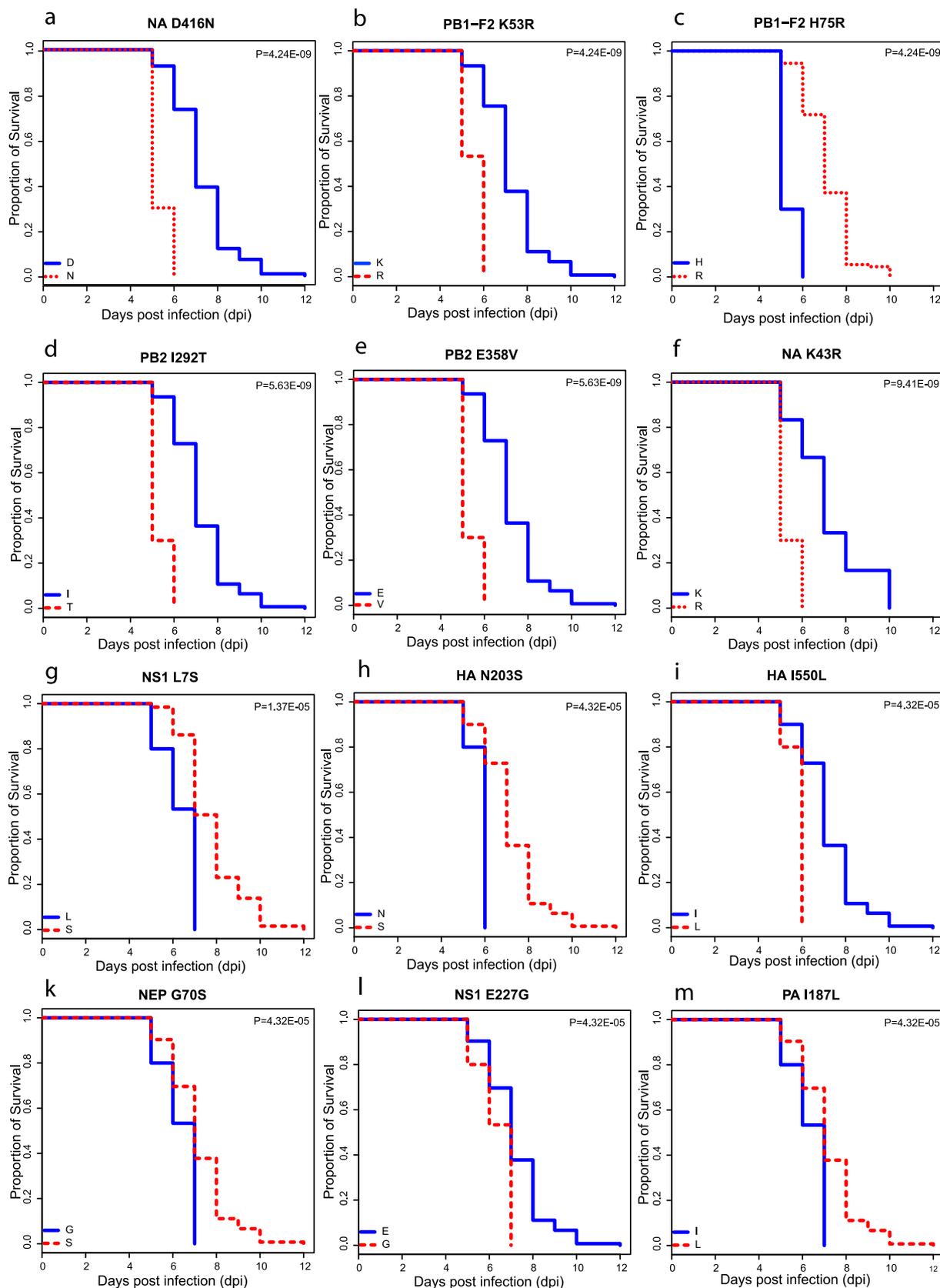


Figure 2 | Polymorphic sites associated with pathogenicity in mice by residue effect. Amino acid substitutions at certain positions affect the pathogenicity of H1N1 IAVs in DBA/2J mice. Sites showing significant residue effect were identified by Cox proportional hazard model after adjusting the host effect. Adjusted p-value (FDR) ≤ 0.01 for residue effect was deemed significant. The variants (a) N416D in NA, (b) K53R and (c) R75H in PB1-F2, (d) I292T and (e) E358V in PB2, (f) K43R in NA, (g) S7L in NS1, (h) S203N and (i) I550L in HA, (k) S70G in NEP, (l) E227G in NS1, and (m) L187I in PA are associated with increased pathogenicity in avian H1N1 IAVs. These positions were based on the sequences used in this study.

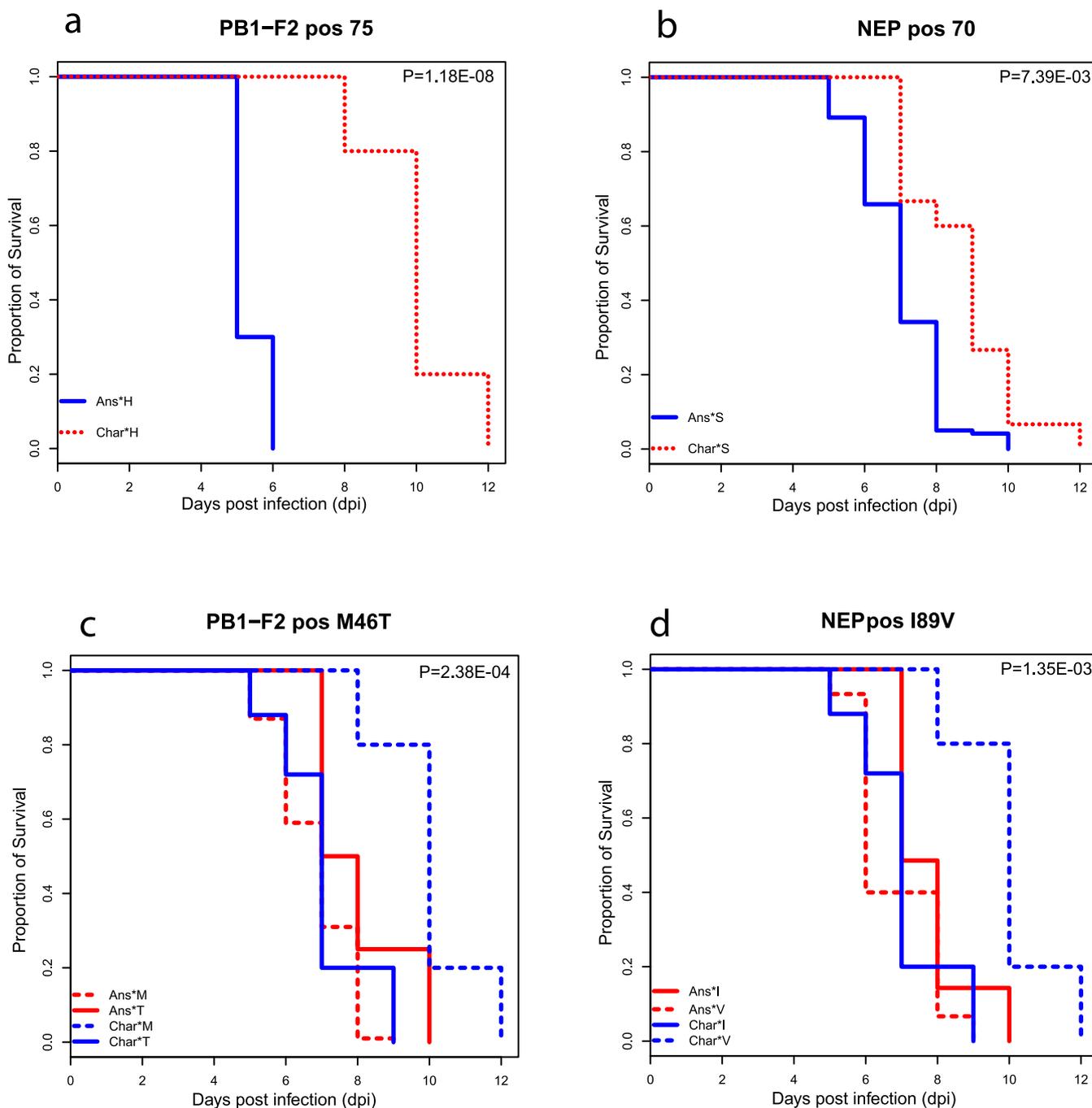


Figure 3 | Polymorphic sites associated with pathogenicity in mice by host origin and host-residue interaction. Amino acid substitutions at certain positions (pos) affect the pathogenicity of H1N1 IAVs of Anseriformes (Ans) or Charadriiformes (Char) origin in DBA/2J mice. Sites showing significant host effect and host-residue interactions were identified by Cox proportional hazard model with host, residue, and host-residue interaction term in the model. Adjusted p-value (FDR) ≤ 0.01 was deemed significant. Two of the polymorphic sites with host effect: (a–b) H75 in PB1-F2, and S70 in NEP are individually associated with increased pathogenicity in Ans IAVs. Two of the polymorphic sites with host-residue interactions: (c–d) T46M in PB1-F2 and I89V in NEP are individually associated with increased pathogenicity in Ans IAVs, and M46T in PB1-F2 and V89I in NEP have the same effect in Char IAVs. These positions were based on the sequences used in this study.

tioned above. Positively charged K73 and R75 residues within a 12-aa subdomain (residues 63–75) are the minimal requirement for a full-length peptide to localize to the mitochondria instead of cytoplasm¹⁸. Here we observed that K73 was conserved in 31 H1N1 IAVs. Although R75H substitution in PB1-F2 of H1N1 IAVs was associated with pathogenicity in mice by residue effect (Fig. 2c and Table 2), localization of this protein was likely perturbed. Strains with variants in position 75 also have coevolved positive residues

in the surrounding subdomain Q69R and K78R, which might compensate for the loss of R75.

For PB1-F2, we also identified seven sites (8, 15, 23, 27, 69, 75, 78) associated with the pathogenicity by host effect (adjusted P (FDR) $\leq 9.62E-03$; Supplementary Table S3). On the basis of host-residue interactions, we identified four polymorphic sites. The residues R26, T46, S83, and N90 showed increased pathogenicity in avian H1N1 IAVs of Charadriiformes origin, whereas, the other four res-



Table 2 | Selected statistically significant polymorphic sites and their potential function or structural insights

Protein	Site	Residues	Effect	Location/Potential Effect
PB2	358	E, V	Residue only	Cap-binding domain (318-483 aa) ³⁷
	649	V, I	Residue only	Host-specific domain (538-678 aa) ³⁹
PB1-F2	199	A, T	Residue and host	Genomic marker between avian and human IAVs ¹⁷
	75	H, L, R	Residue and host	Part of a positively charged subdomain required for the localization of full-length peptide to mitochondria ¹⁸
PA	323	V, I	Host-residue interaction	Compensate for the variation at site 545 to stabilize the molecule
	545	I, V	Residue only	Compensate for the variation at site 323 to stabilize the molecule
HA	187	L, I	Residue and host	N-terminal endonuclease domain of PA (1-196 aa) ⁵⁰
	173	N, T	Residue and host	Sa antigenic site ⁵¹
NS1	203	N, S	Residue only	Sb antigenic site ⁵¹
	7	S, L, T	Residue only	dsRNA binding domain ⁵²
	27	I, M	Residue only	dsRNA binding domain ⁵²
	101*	D, E	Residue only	Interacts with the P13K-inhibitory domain; might act as nuclear export signal and key residue for the NS1-CPSF30 interaction in H5N1 viruses ⁵³
	103*	F, Y	Residue only	Interacts with the P13K-inhibitory domain, increase in IFN antagonism when tested in H5N1 background ⁵⁴
	118*	R, K	Residue only	Allows closer packing of the P13K helix to the NS1 protein
NEP	227	E, G	Residue and host	PDZ-binding motif; genomic marker between avian and human IAVs ¹⁷
	7	S, L, T	Residue only	protease-sensitive N-terminal domain ²¹ ; function unknown
	70	S, G	Residue and host	M1 binding domain ²¹ ; genomic marker between avian and human IAVs ¹⁷
	64*	K, T	Residue only	M1 binding domain ²¹ ; increases the negative charge of the M1-binding domain
	86*	R, I	Residue only	M1 binding domain ²¹ ; increases the negative charge of the M1-binding domain
	88*	K, T	Residue only	M1 binding domain ²¹ ; increases the negative charge of the M1-binding domain

These positions were based on the sequences used in this study.

*The polymorphic sites with underlying allele effect.

idue Q26, M46, F83, and S90 showed increased pathogenicity in those of Anseriformes origin (adjusted P (FDR) $\leq 8.66E-03$; Supplementary Table S4). For instance, the host-residue interaction at position 46 in PB1-F2 indicated that M46T substitution was associated with increased pathogenicity of the viruses of Charadriiformes origin and decreased pathogenicity of viruses of Anseriformes origin (Fig. 3c).

We found that five of 45 polymorphic sites in the PA protein (L187I, R269K, I348L, S388G, and I545V) were associated with pathogenicity by residue variations (adjusted P (FDR) $\leq 5.17E-03$; Supplementary Table S2). The site, 187, was also associated with the pathogenicity by host effect (adjusted P (FDR) = 7.39E-03; Table 2, Supplementary Table S3). The L187I substitution was observed in influenza viruses of shorebird origin that were more pathogenic in mice (Supplementary Fig. S4), but the viruses of gull or Anseriformes origin did not carry this substitution. According to host-residue interaction, D272, I323 and P400 were associated with increased pathogenicity of avian H1N1 IAVs of Anseriformes origin; conversely, E272, V323, and Q400 appeared to help retain pathogenicity of avian H1N1 IAVs of Charadriiformes origin (adjusted P (FDR) = 8.83E-03; Supplementary Table S4).

The recently discovered PA-X protein is a product of the X-ORF within the PA-ORF and resulted from +1 ribosomal frameshifting, the 61-aa long C terminus of which fuses with the N terminus of PA³. We detected 12 polymorphic sites in this region, none of which were associated with pathogenicity by residue or host-origin effect (Supplementary Fig. S5). However, we identified four sites with substitutions associated with the pathogenicity by host-residue interactions. The L16, V21, S23, and P59 were associated with increased pathogenicity of viruses of Charadriiformes origin, and S16, A21, L23, and Q59 were associated with increased pathogenicity of viruses of Anseriformes origin (Supplementary Table S4; adjusted P (FDR) was 9.02E-03 for position 21, and 8.83E-03 for other sites).

Surface glycoproteins. We detected 49 variants in HA of North American avian H1N1 IAVs (Supplementary Fig. S6). Two variants (S203N and I550L) were associated with pathogenicity by residue effect (Fig. 2h, i; adjusted P (FDR) = 4.32E-05; Supplementary Table S2). The variant S203N was located in the Sb antigenic site (Table 2), possibly playing a role in host immune response (Fig. 4). We identified two polymorphic sites (173 and 295) affecting the pathogenicity of viruses in mice by host effect (adjusted P (FDR) was 9.96E-03 and 9.50E-03, respectively; Supplementary Table S3). Position 173, which is located in the head of the HA protein, is part of the Sa antigenic site (Table 2) and plays a role in host immune response (Fig. 4).

We detected 44 variants in the NA of avian H1N1 IAVs (Supplementary Fig. S8). Two variants (K43R and D416N) were associated with pathogenicity in mice by residue effect (Figs. 2a, f; adjusted P (FDR) was 4.24E-09 for site 416 and 9.41E-09 for site 43; Supplementary Table S2). One of the most significant variant, D416N, is located at the interface between the monomers (Supplementary Fig. S17b).

Matrix and nonstructural proteins. We identified two variants in M1 and nine in M2 proteins of North American avian H1N1 IAVs (Supplementary Fig. S9a, b). The only variant associated with pathogenicity in mice by residue effect, R18K, was identified in the M2 protein (adjusted P (FDR) = 4.23E-04; Supplementary Table S2).

Based on their sequence homology, the NS gene sequences are grouped as allele A (swine/human-like) or allele B (avian-like)¹⁹. The avian H1N1 IAVs in this study belonged to both allele A ($n=17$) and allele B ($n=14$). Therefore, the number of polymorphic sites in the NS gene products was higher due to the underlying allele effect.

Of the 80 polymorphic sites observed in NS1, 60 sites (75%) were initially associated with the pathogenicity of viruses in mice by res-

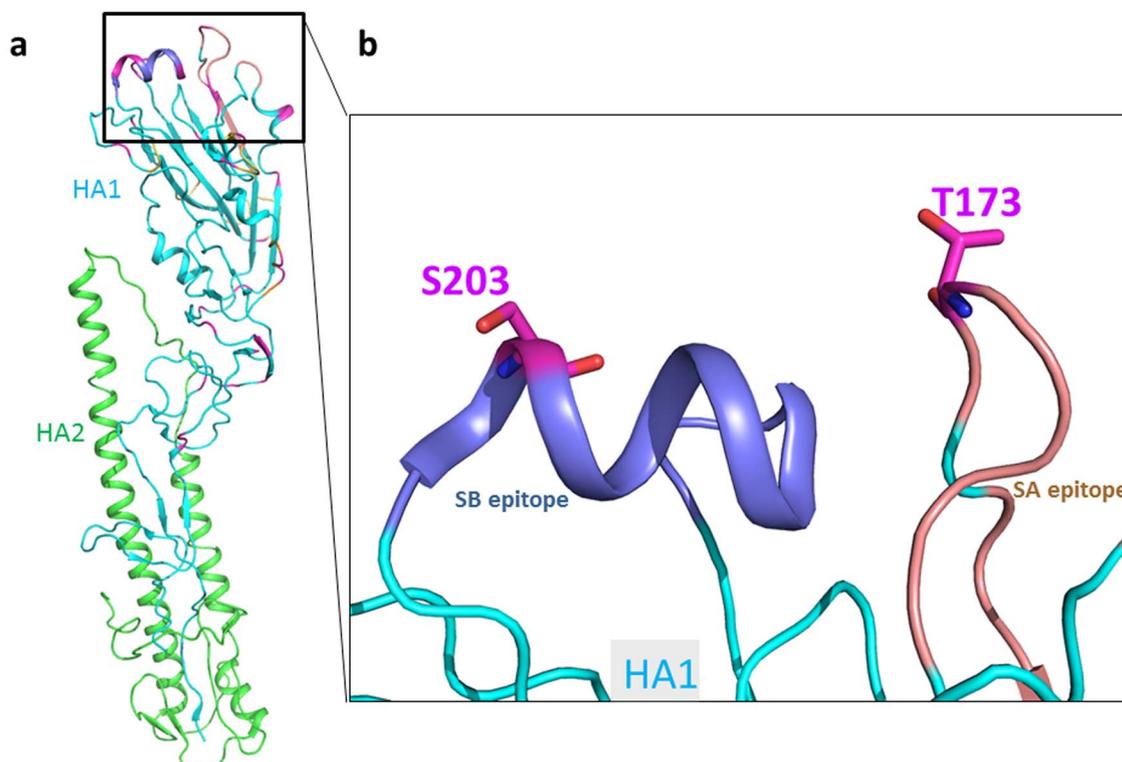


Figure 4 | The HA variants associated with pathogenicity of avian H1N1 IAVs by residue and host effect. (a) Ribbon model of the HA1 (teal) and HA2 (green) dimer in a single HA molecule from *A/mallard/Alberta/35/1976* [PDB: 2WRH]. Antigenic sites Sa (peach), Ca (yellow), Sb (dark blue), and Cb (orange) and locations of variants (magenta) are indicated. (b) The variants S203 and T173 in HA of avian H1N1 IAVs are located in the head of HA, as part of the Sb (dark blue) and Sa (peach) antigenic sites, respectively. These variants play a role in host immune response. These positions were based on the sequences used in this study.

idue effect (adjusted P (FDR) $\leq 4.77\text{E-}03$; Supplementary Table S2). After adjusting the underlying allele effect, 4 polymorphic residues (S7L, T7L, I27M, and E227G) in NS1 remained significant by residue effect (adjusted P (FDR) values are $1.37\text{E-}05$, $4.77\text{E-}03$, $1.87\text{E-}03$, and $4.32\text{E-}05$, respectively; Supplementary Table S2). The host effect was detected at only one site (227) (adjusted P (FDR) = $7.39\text{E-}03$; Supplementary Table S3). On the basis of host-residue interactions, we observed that S206 was associated with increased pathogenicity of H1N1 IAVs of Charadriiformes origin, and C206 was associated with increased pathogenicity in those of Anseriformes origin (adjusted P (FDR) = $4.19\text{E-}03$; Supplementary Table S4).

Several of the polymorphic sites identified in association with pathogenicity by residue effect in NS1, but showing underlying allele effect, were found to have structural and functional importance. Modeling two of those variants, D101E and F103Y, low-pathogenic strains contain residues D101 and F103, but only the former interacts with the phosphoinositide 3-kinase (PI3K)-inhibitory domain at R587 (Table 2). Pathogenic viruses contain E101 and Y103, making it more plausible that both residues would be involved in the interaction (Fig. 5a). D101E would retain a similar salt bridge with R587, but the additional hydroxyl on Y103 would facilitate further interaction. The outcome would be a stronger interaction between NS1 and the PI3K-inhibitory domain.

Among the 22 polymorphic sites on NEP (Supplementary Fig. S11), we identified 18 residues that were associated with pathogenicity by residue effect (adjusted P (FDR) $\leq 4.77\text{E-}03$; Supplementary Table S2); host effect on pathogenicity occurred only at site 70 (adjusted P (FDR) = $7.39\text{E-}03$; Supplementary Table S3). After adjusting the underlying allele effect, we still found 3 polymorphic residues in NEP by residue effect, S7L, T7L, and S70G (adjusted P (FDR) values are $1.37\text{E-}05$, $4.77\text{E-}03$, and $4.32\text{E-}05$, respectively;

Supplementary Table S2). On the basis of host-residue interactions, we observed an increased pathogenicity of avian H1N1 IAVs of Anseriformes origin in mice due to I89V substitution in NEP (adjusted P (FDR) = $1.35\text{E-}03$; Fig. 3d and Supplementary Table S4). The several statistically associated variants with underlying allele effect, K64T, R86L, and K88T, were also present on the M1-binding surface of more pathogenic strains increasing the negative charge of the M1-binding domain (wild-type charge at pH 7.00 = -1.8 ; variants charge at pH 7.00 = -5.1) (Fig. 5b). On the basis of this finding, we speculated that the increased negative charge on the M1-binding surface of the more pathogenic strains provides an additional advantage for early infection due to tighter interaction of NEP and M1. Because M1 is not present in large quantities until later stages of infection, the tighter binding between NEP and M1 may cause early export of genomic ribonucleoproteins from the nucleus²⁰.

Interestingly, the variant E227G in NS1 and S70G in NEP, caused by the same nucleotide substitution c.A680G in NS segment, showed both host and residue effect (Fig. 2l and Fig. 3d). The E227G/S70G was observed in three pathogenic viruses of Charadriiformes origin (*A/shorebird/DE/300/2009*, *A/shorebird/DE/324/2009*, and *A/gull/DE/428/2009*) with 100% mortality by 7 dpi in mice (Supplementary Table S1). In NS1, the E227G alters the PDZ-binding motif from ESEV to GSEV (Table 2). Species with ESEV contain a di-Arg (RRVESEV) in the upstream of PDZ-binding motif, and those with GSEV do not (RTIGSEV). We detected 13 viruses with an RTIESEV motif, nine of which belong to the viruses of lower pathogenicity. In contrast, the RTIGSEV motif was found only in H1N1 IAVs of Charadriiformes origin with greater pathogenicity. In NEP, S70 is avian like and G70 is mammalian-like residue that is observed in swine, human, and the 2009 pandemic H1N1 isolates¹⁷. It is located in NEP-M1 binding domain on NEP (residues 54-121)²¹.

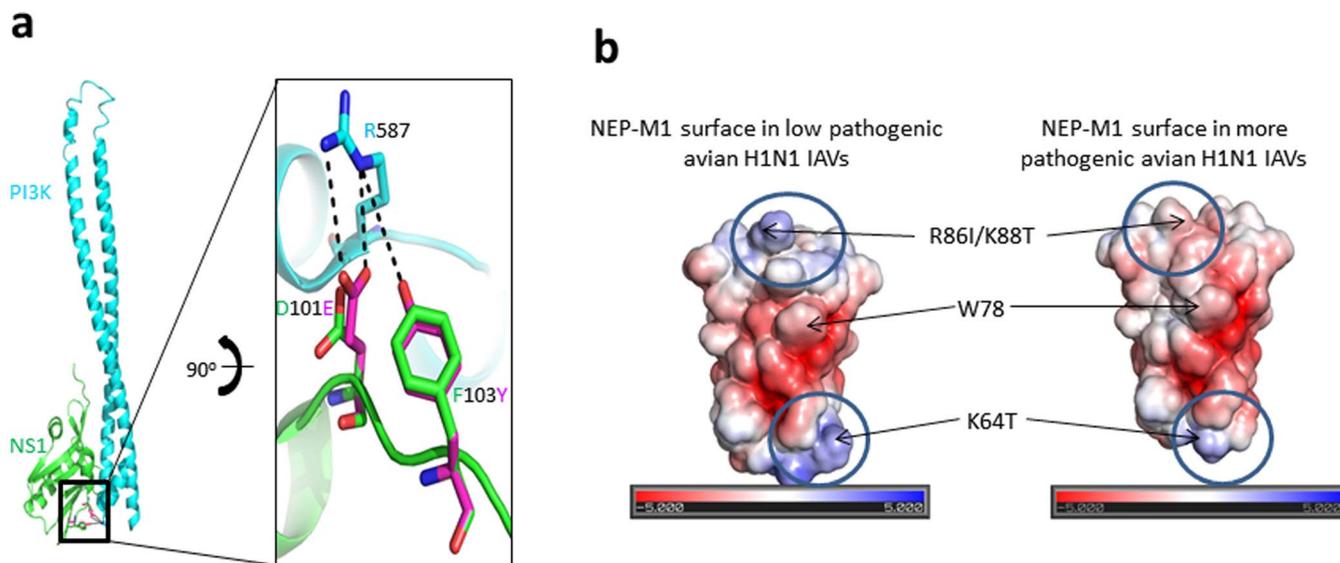


Figure 5 | NS1 and NEP variants associated with pathogenicity by residue effect. (a) Ribbon model of NS1 (green) in which S103 is replaced with F103, and NS1 is bound to the PI3K inhibitory domain (teal) [Bovine PDB: 3L4Q] are labeled. The predicted interaction (black dots $<5 \text{ \AA}$) between variant residues E101 and Y103 of NS1 protein with R587 (numbering is based on human) from the PI3K inhibitory domain. Low-pathogenic strains contain residue D101 and F103 in which D101 interacts only with the PI3K inhibitory domain R587. Pathogenic viruses contain E101 and Y103, and both residues are involved in interactions. The outcome is a stronger interaction between NS1 and the PI3K inhibitory domain. (b) NEP variants increase the negative charge of the M1-binding domain in more pathogenic viruses. Electrostatic surface potential representation of the NEP/M1-binding surface [PDB: 1PD3]; native (left image) and with three variant changes (R86I, K88T, and K64T) (right image). The surface color indicates negative charge (red), positive charge (blue), and neutral (white). The extent of the charge is indicated below each image in units of kT/e . These positions were based on the sequences used in this study.

Discussion

The pathogenicity of IAVs is a polygenic trait that results from combinations of genetic variants in different gene segments of the virus. Understanding the genomic dynamics of IAVs is an important step in risk assessment, in terms of determining the virus' virulence, transmissibility, and pandemic potential. Previously, we introduced a measurement named "pathogenicity score" to indicate the overall effect of each virus on the infected host¹². Here we applied the Cox proportional hazard model to the survival data obtained from individual IAV infected mice in order to identify crucial natural polymorphisms in the genome of wild-type North American avian H1N1 IAVs in terms of their pathogenicity in a mammalian model. Although the same statistical model has been widely used in human genomics, to our knowledge, it's the first time that this method is applied in the virus pathogenicity studies. This method enables us to directly estimate the relationship between covariates such as genetic variants, hosts, and the time of death. The Cox proportional hazard model also allows us to assess the individual contribution of covariates to the model.

Our results show that at the deduced proteome level, the pathogenicity of the avian H1N1 viruses in DBA/2J mice is associated with 32 polymorphic sites by residue effect (after adjusting the underlying allele effect by NS segment), 23 polymorphic sites by host effect, and 14 polymorphic sites by host-residue interactions. Overall, pathogenicity of avian H1N1 IAVs in mice was mainly associated with polymerase complex (PB2, PB1, PB1-F2, PA, PA-X) and NS gene products (NS1 and NEP). The highest number of polymorphic sites ($n=73$) in the NS gene sequences was due to the circulation of two alleles (Allele A and B) in North American IAVs of various subtypes^{19,22}. The viruses in our study fell into both allele groups, but the avian-like allele (Allele B) appeared to be more associated with pathogenicity in mice. Thus, the effect of underlying allele of NS segment played a significant role on the pathogenicity of avian H1N1 IAVs in mice.

Modeling the variants associated with pathogenicity using available crystal structures helped us understand the functional and struc-

tural importance of those sites. For example, some significant variants associated with pathogenicity, E358V and V649I, were located in the cap-binding and host-specific domains of PB2, respectively. The substitutions A199S and V667I in PB2 were previously identified in association with enhanced transmission to humans^{17,23}. Here, we detected A199T substitution associated with increased pathogenicity in mice, which was observed in five H1N1 IAVs of Charadriiformes origin (A/shorebird/DE/300/2009, A/shorebird/DE/324/2009, A/gull/DE/428/2009, A/shorebird/DE/170/2009, and A/shorebird/DE/318/2009). Threonine is biochemically more similar to serine than to alanine. Therefore, H1N1 viruses of Charadriiformes origin might use A199T substitution as a step toward adapting to mammals. Given that V667I is observed only in A/pintail/ALB/210/2002, it could not be detected by statistical analysis due to the lack of statistical power.

The PB1 of avian IAVs mostly carry N375 although S375 and T375 can also be found at lower proportion (18% and 13%, respectively)²⁴. The S375 variant was found in the PB1 of 1918 pandemic H1N1, and human H1N1, H2N2 and H3N2 IAVs²⁴. We detected host-residue interaction at position 375 in PB1 that might indicate the importance of N375S substitution in terms of retaining the pathogenicity of viruses originated from different bird species. All 31 North American avian H1N1 IAVs possessed a full-length PB1-F2. The C-terminal region of the protein mediates mitochondrial localization^{18,25} by interacting with two mitochondrial proteins, ANT3 and VDAC1, to induce apoptosis²⁶. Our analysis showed that one of those sites, 75, found in the 12-aa subdomain of C-terminal of PB1-F2 that was shown as minimal requirement for mitochondrial localization of the protein^{18,25} was associated with pathogenicity by residue and host effect.

The effector domain of NS1 (residues 74-230) interacts with many host proteins to enhance viral mRNA translation²⁷, deregulate cellular mRNA processing²⁸, inhibit double-stranded RNA-activated protein kinase²⁹ and activate PI3K signaling³⁰. Most of the polymorphic residues associated with pathogenicity were in the effector



domain of NS1 of the avian H1N1 IAVs examined. Some were structurally critical for the NS1–PI3K inhibitory domain. Four residues at the C terminus of NS1 are known to interact with PDZ-binding proteins and to alter the host's response. The avian-like PDZ-binding domain (ESEV or EPEV) increases the virulence of the virus in a mouse model³¹. Twenty-eight of 31 avian H1N1 viruses carry an ESEV PDZ-binding domain that explains their pathogenicity in mouse models. However, we detected the GSEV PDZ-binding motif in three Charadriiformes isolates (A/shorebird/DE/300/2009, A/shorebird/DE/324/2009, and A/gull/DE/428/2009) that were previously categorized in the most pathogenic or moderately pathogenic virus groups¹². A crystal structure of the PDZ domain with a bound ESEV peptide shows that the E227 residue on NS1 forming a salt bridge with R16 on the PDZ peptide³². The R227 (RSEV motif) seen in some human influenza viruses²² would not form this interaction but rather it would be electrostatically unfavorable. G227 would not form a salt bridge similar to that of E227–R16, nor would it cause electrostatically unfavorable interactions. Therefore, GSEV would most likely bind the same PDZ domains that ESEV binds, except with less stabilizing interactions. We speculate that two adjacent arginines at residues 224 and 225 are complemented by the acidic ESEV motif, which also allows interaction with a basic charged PDZ-binding domain. In the GSEV motif, the presence of the two adjacent arginines could shift the C terminus to be more basic and repel the peptide from the PDZ domain. Thus, these two residues provide an advantage for the interactions of ESEV with PDZ-binding domains, while arginine-threonine (RT) provides an advantage to GSEV. This hypothesis is supported by RT occurring upstream of the GSEV in more than 90% of the sequences in the publicly available database. The distribution of GSEV motif in various subtypes, hosts and geographical regions in the full-length publicly available NS1 sequences is given in the Supplementary Information (Supplementary Tables 6, 7). Overall, NS appears to play a key role in the pathogenicity of avian H1N1 IAVs in mammalian models.

We observed host-originated differences (Anseriformes versus Charadriiformes) in terms of the distribution and number of polymorphic sites in the full genome. According to the overall proteotype profiles, the six North American avian H1N1 IAVs isolated from Charadriiformes in 2009 displayed host-specific genetic differences such as shorebird versus gull, but the 25 influenza viruses isolated from Anseriformes (mallards, northern shovellers, and pintails) over 24 years (1984–2008) did not. During the surveillance studies conducted in Delaware Bay, H1N1 subtype was not detected in Charadriiformes until 2009 (personal communication with Scott Krauss, St. Jude). The six H1N1 isolates of Charadriiformes origin in the St. Jude influenza repository differ from other avian H1N1 viruses at 10 positions in their HAs, but they do not differ in their NAs. We detected 11 positions associated with pathogenicity by residue effect which were found only in the polymerase genes of five H1N1 IAVs of Charadriiformes origin (A/shorebird/DE/300/2009, A/shorebird/DE/324/2009, A/gull/DE/428/2009, A/shorebird/DE/170/2009, and A/shorebird/DE/318/2009). We also detected two human-like residues (D382 and N409) already circulating in the PA gene pool of North American avian H1N1 IAVs of mainly Charadriiformes origin (human-like N409 in A/shorebird/DE/300/2009, A/shorebird/DE/324/2009, A/shorebird/DE/170/2009, and A/shorebird/DE/318/2009; and human-like D382 in A/shorebird/DE/274/2009). Although our sample size for H1N1 IAVs of Charadriiformes origin is too limited to make a solid conclusion, the unique substitutions may provide an additional advantage to the polymerase complex of avian H1N1 viruses of Charadriiformes origin to replicate more efficiently and to transmit in different host systems such as mammals.

Because genetic changes occur in multiple genes, it is difficult to identify exact residues and/or the combination of amino acid substitutions that plays a role in adaptation, transmission, or virulence of

the virus. However, combining our previous findings on the disease-causing potential of avian H1N1 IAVs with those from this study serves as an initial attempt to reveal the pathogenic genomic signatures in a naturally circulating gene pool. With this novel approach, we highlighted the importance of some residues in each protein that might play a role in the pathogenicity of the virus by altering protein-protein interactions. Variations at individual residues appear to be host- and virus strain-specific; however, using this sort of approach, we might be able to pinpoint key locations on the proteins that are altered by selective pressure. Our analyses also emphasized the importance of genetic variants occurring in viruses of different bird populations. Overall, the statistical approach used to understand the importance of amino acid substitutions, the selective pressure on the different sites of the genome, and the simulation of the polymorphic sites through structural modeling helped us understand the significance of certain genomic variations on the pathogenicity of avian H1N1 IAVs in mammalian models.

This comprehensive analysis as an initial attempt advances the field of influenza evolution and host adaptation by not only identifying polymorphic sites associated with pathogenicity, but also investigating where those residues are located on the proteins and their structural and functional contribution to pathogenicity. In the long run, in conjunction with broader surveillance efforts, the ultimate goal is to deduce the virulence of wild-type avian IAVs and other emerging zoonotic pathogens in mammals based on their genomic information.

Methods

Viruses. We analyzed 31 North American avian H1N1 IAVs (Supplementary Table S1) to investigate the genetic relatedness among the viruses and gain understanding about how genomic information of wild-type viruses determines pathogenicity. All viruses were obtained from the St. Jude influenza repository, with the exception of two viruses (A/mallard/OH/4809-9/2008 and A/mallard/MO/466554-14/2007) that were kindly provided by Dennis Senne (U. S. Department of Agriculture National Veterinary Services Laboratories, Diagnostic Virology Section, USDA-APHIS, Ames, IA), Tom DeLiberto and Seth Swafford (both of the U.S. Department of Agriculture Animal and Plant Health Inspection Service, Wildlife Services, National Wildlife Disease Program, Fort Collins, CO). Viruses were minimally propagated in embryonated chicken eggs as previously described¹². Viral RNA was extracted using QIAamp® Viral RNA Kit (Qiagen, Gaithersburg, MD) per the manufacturer's instructions.

Whole-genome sequencing. The genomes of 31 avian H1N1 viruses (Supplementary Table S1) were sequenced using high-throughput (Illumina) and traditional (Sanger) platforms. Briefly, cDNA libraries were prepared using a SuperScript III One-Step RT-PCR Platinum Taq HiFi kit (Invitrogen, Grand Island, NY). Then, cDNA was sheared, end-repaired, and a poly-A tail was added. Finally, adapters were ligated before the index sequences were inserted. High-throughput sequencing was done on an Illumina Genome Analyzer IIx system (Illumina, San Diego, CA). For Sanger sequencing, each segment was amplified using previously described gene-specific primers³³. For efficient amplification of HA gene sequences from H1N1 IAVs, modified versions of previously described H1 primers were used (forward: 5' AGCAAAAGCAGGGGAAATTCAAAATC 3' and reverse: 5' AGTAGAAACAAGGGTGTTTTCCACA 3'). Each gene segment was amplified using One-Step RT-PCR Kit (Qiagen), and the amplicons were purified by QIAquick gel extraction kit (Qiagen) per the manufacturer's instructions. Each amplicon was sequenced with Big Dye® Terminator (v3.1) Chemistry on Applied Biosystems 3730XL DNA Analyzers (Life Technologies, Carlsbad, CA). Sequence reads were assembled using DNASTAR Lasergene 9 Core Suite Seqman (DNASTAR, Inc., Madison, WI). The consensus sequences were aligned using MEGA 5.0 and checked for the genetic variants.

Proteotyping. The protein-coding sequences of each gene segment were aligned using ClustalW multiple alignment on the BioEdit Sequence Alignment Editor version 7.0.9.0. Conserved positions in each alignment were removed, and the amino acids were color coded at the variable sites using a seqplot program (gcpplot and residueplot) written in Perl. Similar amino acids were grouped by color: aliphatic amino acids, shades of brown and orange; positively charged residues, shades of blue; negatively charged ones, shades of red; polar uncharged ones, shades of green; aromatic ones, shades of purple; and glycine and cysteine, shades of yellow. Gaps and unknown residues were indicated in dark gray. The individual color codes used for proteotype profile were assigned as follows: A: medium orange, C: yellow, D: red, E: pink, F: dark violet, G: yellow-green, H: pure violet, I: light orange, K: light blue, L: dark brown, M: medium brown, N: dark lime green, P: dark orange, Q: light green, R: dark blue, S: lime green, T: dark cyan, V: light brown, W: light violet, X: light gray, Y:



maroon and gap(-): dark gray. For each strain, all protein sequences were combined (PB2+PB1+PB1-F2+, etc.) and clustered using ClustalW, so that viruses would be ordered according to their pathogenicity (highest to lowest). Then the viruses were clustered based on the concatenated nucleotide sequences of their 12 expressed gene products. Aligned sequences were manually curated for the presence of previously described pathogenicity, mammalian adaptation, and drug-resistance markers.

Statistical analyses. We investigated the effect of single-residue variations on the proteome of 31 North American H1N1 IAVs in association with their pathogenicity previously identified in DBA/2J mice¹². Additionally, we investigated the effect of host origin of virus strains and the host-residue interactions in relation to the pathogenicity. Survival data was used to examine the pathogenicity of each selected virus in the infected mice (n=150 mice; 5 mice/virus strain). A/green-winged teal/LA/Sg-00090/2007 caused death in only 1 mouse 11 dpi; thus, it was defined as an outlier and excluded from the data analysis.

More specifically, the Cox proportional hazard model was used to evaluate the association between the mouse survival time after exposure to the virus and amino acid residues at each position, host origin for viruses (Anseriformes or Charadriiformes), the interaction between amino acid residue and virus host, simultaneously. The coefficient for each explanatory variable in the Cox model have been adjusted for confounding by the other explanatory variables, i.e. in effect when all other explanatory variables are held constant. For amino acid residue, we required that there were at least two groups of residues at each position and there were at least two viruses at each amino acid group. We then sorted the amino acid residues at each polymorphic site in alphabetical order; the first one was used as reference. For host effect, Anseriformes were chosen as the reference.

The Cox proportional hazard model is a semiparametric model that uses maximum-likelihood (ML) estimation to estimate the coefficient for each parameter while controlling the confounding effects of other included variables. The likelihood is the product of several likelihoods, one for each death time. Once the ML estimation is complete, the Wald test is used to determine the p-value for the coefficient²⁴. Wald test uses z-statistics obtained by dividing the estimated coefficient by its standard error then assuming that this quantity is approximately a standard normal. The positive sign of the coefficient means that the hazard (risk of death) is higher for a mouse exposed to the virus with higher value of that parameter. The positive coefficient for a residue showed that the non-reference residue was more pathogenic after adjusting the host effect and vice versa. The positive coefficient for a host showed that H1N1 viruses of Charadriiformes origin were more pathogenic after adjusting the residue effect and vice versa. The interaction term showed whether the residue effect on survival differed for the two hosts; for instance, a substitution at a site is associated with increased pathogenicity in one type of host origin and decreased pathogenicity in another. The p-value for each factor was adjusted using Benjamini & Hochberg method ("BH" or its alias "FDR")³⁵. The adjusted p-value of 0.01 or less was deemed significant. Significant residues that conferred more pathogenicity but appeared in the least pathogenic strain, A/green-winged teal/LA/Sg-00090/2007, were removed. All the analyses were done by survival package in R 2.15.1 (<http://www.r-project.org/>).

Structural modeling and interpretation of H1N1 viral variants. Structures that overlapped the following proteins were downloaded from the PDB³⁶: PB2 [PDB: 2VQZ³⁷ and 4ENI³⁸ for cap-binding domain (318–483 aa); and 2VY8³⁹ for host-specific domain (538–678 aa)]; PB1 [PDB: 3CM8, 2ZTT⁴⁰], PA [PDB: 2ZNL⁴¹], HA [PDB: 2WRH⁴²], NP [PDB: 3RO5⁴³], NA [PDB: 2HTY⁴⁴ and 3B7E⁴⁵], M1 [PDB: 1EA3⁴⁶], M2 [PDB: 2H95⁴⁷], NS1 [PDB: 3L4Q⁴⁸, 3D6R⁴⁹ and 3PDV³²], and NEP [PDB: 1PD3²¹]. NS1/PI3K-interface residues were identified with LigPlot +. Figures, mutations, and analyses were performed using PyMol. Surface potentials were estimated using the APBS plugin in PyMol. Linear electrostatic charge was estimated using default options in the protcalc webserver (<http://protcalc.sourceforge.net/>). The location of the variants was visually inspected on available structures for potential functional impact.

Data Access. The sequences for the whole genome of 31 North American avian H1N1 viruses sequenced in this study can be accessed from the NCBI GenBank under accession numbers KF424015–KF424262 (Supplementary Table S1).

- Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. Evolution and ecology of influenza A viruses. *Microbiol. Rev.* **56**, 152–179 (1992).
- Chen, W. *et al.* A novel influenza A virus mitochondrial protein that induces cell death. *Nat. Med.* **7**, 1306–1312, doi:10.1038/nm1201-1306 (2001).
- Jagger, B. W. *et al.* An overlapping protein-coding region in influenza A virus segment 3 modulates the host response. *Science* **337**, 199–204, doi:10.1126/science.1222213 (2012).
- Steinhauer, D. A. & Skehel, J. J. Genetics of influenza viruses. *Annu. Rev. Genet.* **36**, 305–332, doi:10.1146/annurev.genet.36.052402.152757 (2002).
- Wise, H. M. *et al.* A complicated message: Identification of a novel PB1-related protein translated from influenza A virus segment 2 mRNA. *J. Virol.* **83**, 8021–8031, doi:10.1128/JVI.00826-09 (2009).
- Fouchier, R. A. *et al.* Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. *J. Virol.* **79**, 2814–2822, doi:10.1128/JVI.79.5.2814-2822.2005 (2005).
- Tong, S. *et al.* A distinct lineage of influenza A virus from bats. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 4269–4274, doi:10.1073/pnas.1116200109 (2012).
- Tong, S. *et al.* New world bats harbor diverse influenza A viruses. *PLoS Pathog.* **9**, e1003657, doi:10.1371/journal.ppat.1003657 (2013).
- Zhu, X. *et al.* Crystal structures of two subtype N10 neuraminidase-like proteins from bat influenza A viruses reveal a diverged putative active site. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 18903–18908, doi:10.1073/pnas.1212579109 (2012).
- Kanegae, Y., Sugita, S., Shorridge, K. F., Yoshioka, Y. & Nerome, K. Origin and evolutionary pathways of the H1 hemagglutinin gene of avian, swine and human influenza viruses: cocirculation of two distinct lineages of swine virus. *Arch. Virol.* **134**, 17–28 (1994).
- Reid, A. H., Fanning, T. G., Hultin, J. V. & Taubenberger, J. K. Origin and evolution of the 1918 "Spanish" influenza virus hemagglutinin gene. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 1651–1656 (1999).
- Koçer, Z. A., Krauss, S., Stallknecht, D. E., Reh, J. E. & Webster, R. G. The potential of avian H1N1 influenza A viruses to replicate and cause disease in mammalian models. *PLoS One* **7**, e41609, doi:10.1371/journal.pone.0041609 (2012).
- Subbarao, E. K., London, W. & Murphy, B. R. A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. *J. Virol.* **67**, 1761–1764 (1993).
- Kageyama, T. *et al.* Genetic analysis of novel avian A(H7N9) influenza viruses isolated from patients in China, February to April 2013. *Euro Surveill.* **18**, 20453 (2013).
- Han, J. Y. *et al.* A genome-wide association study of survival in small-cell lung cancer patients treated with irinotecan plus cisplatin chemotherapy. *Pharmacogenomics J.* **14**, 20–27, doi:10.1038/tpj.2013.7 (2014).
- Wu, C. *et al.* Genome-wide association study of survival in patients with pancreatic adenocarcinoma. *Gut* **63**, 152–160, doi:10.1136/gutjnl-2012-303477 (2014).
- Chen, G. W. *et al.* Genomic signatures of human versus avian influenza A viruses. *Emerg. Infect. Dis.* **12**, 1353–1360, doi:10.3201/eid1209.060276 (2006).
- Yamada, H., Chounan, R., Higashi, Y., Kurihara, N. & Kido, H. Mitochondrial targeting sequence of the influenza A virus PB1-F2 protein and its function in mitochondria. *FEBS Lett.* **578**, 331–336, doi:10.1016/j.febslet.2004.11.017 (2004).
- Treanor, J. J., Snyder, M. H., London, W. T. & Murphy, B. R. The B allele of the NS gene of avian influenza viruses, but not the A allele, attenuates a human influenza A virus for squirrel monkeys. *Virology* **171**, 1–9 (1989).
- Paterson, D. & Fodor, E. Emerging roles for the influenza A virus nuclear export protein (NEP). *PLoS Pathog.* **8**, e1003019, doi:10.1371/journal.ppat.1003019 (2012).
- Akarsu, H. *et al.* Crystal structure of the M1 protein-binding domain of the influenza A virus nuclear export protein (NEP/NS2). *EMBO J.* **22**, 4646–4655, doi:10.1093/emboj/cdg449 (2003).
- Obenauer, J. C. *et al.* Large-scale sequence analysis of avian influenza isolates. *Science* **311**, 1576–1580, doi:10.1126/science.1121586 (2006).
- Yang, J. R. *et al.* New variants and age shift to high fatality groups contribute to severe successive waves in the 2009 influenza pandemic in Taiwan. *PLoS One* **6**, e28288, doi:10.1371/journal.pone.0028288 (2011).
- Taubenberger, J. K. *et al.* Characterization of the 1918 influenza virus polymerase genes. *Nature* **437**, 889–893, doi:10.1038/nature04230 (2005).
- Gibbs, J. S., Malide, D., Hornung, F., Bennis, J. R. & Yewdell, J. W. The influenza A virus PB1-F2 protein targets the inner mitochondrial membrane via a predicted basic amphiphatic helix that disrupts mitochondrial function. *J. Virol.* **77**, 7214–7224 (2003).
- Zamarin, D., Garcia-Sastre, A., Xiao, X., Wang, R. & Palese, P. Influenza virus PB1-F2 protein induces cell death through mitochondrial ANT3 and VDAC1. *PLoS Pathog.* **1**, e4, doi:10.1371/journal.ppat.0010004 (2005).
- Burgui, I., Aragon, T., Ortin, J. & Nieto, A. PABP1 and eIF4G1 associate with influenza virus NS1 protein in viral mRNA translation initiation complexes. *J. Gen. Virol.* **84**, 3263–3274 (2003).
- Noah, D. L., Twu, K. Y. & Krug, R. M. Cellular antiviral responses against influenza A virus are countered at the posttranscriptional level by the viral NS1A protein via its binding to a cellular protein required for the 3' end processing of cellular pre-mRNAs. *Virology* **307**, 386–395 (2003).
- Li, S., Min, J. Y., Krug, R. M. & Sen, G. C. Binding of the influenza A virus NS1 protein to PKR mediates the inhibition of its activation by either PACT or double-stranded RNA. *Virology* **349**, 13–21, doi:10.1016/j.virol.2006.01.005 (2006).
- Hale, B. G., Jackson, D., Chen, Y. H., Lamb, R. A. & Randall, R. E. Influenza A virus NS1 protein binds p85beta and activates phosphatidylinositol-3-kinase signaling. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 14194–14199, doi:10.1073/pnas.0606109103 (2006).
- Jackson, D., Hossain, M. J., Hickman, D., Perez, D. R. & Lamb, R. A. A new influenza virus virulence determinant: the NS1 protein four C-terminal residues modulate pathogenicity. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 4381–4386, doi:10.1073/pnas.0800482105 (2008).
- Yu, J. *et al.* PDlim2 selectively interacts with the PDZ binding motif of highly pathogenic avian H5N1 influenza A virus NS1. *PLoS One* **6**, e19511, doi:10.1371/journal.pone.0019511 (2011).
- Hoffmann, E., Stech, J., Guan, Y., Webster, R. G. & Perez, D. R. Universal primer set for the full-length amplification of all influenza A viruses. *Arch. Virol.* **146**, 2275–2289 (2001).



34. Kleinbaum, D. & Klein, M. *Survival Analysis: A self-learning text*. 2nd edition, Springer Publishers New York Inc (2005).
35. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289–300 (1995).
36. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
37. Guilligay, D. *et al.* The structural basis for cap binding by influenza virus polymerase subunit PB2. *Nat. Struct. Mol. Biol.* **15**, 500–506, doi:10.1038/nsmb.1421 (2008).
38. Liu, Y. *et al.* Structural and functional characterization of K339T substitution identified in the PB2 subunit cap-binding pocket of influenza A virus. *J. Biol. Chem.* **288**, 11013–11023, doi:10.1074/jbc.M112.392878 (2013).
39. Tarendeau, F. *et al.* Host determinant residue lysine 627 lies on the surface of a discrete, folded domain of influenza virus polymerase PB2 subunit. *PLoS Pathog.* **4**, e1000136, doi:10.1371/journal.ppat.1000136 (2008).
40. Sugiyama, K. *et al.* Structural insight into the essential PB1–PB2 subunit contact of the influenza virus RNA polymerase. *EMBO J.* **28**, 1803–1811, doi:10.1038/emboj.2009.138 (2009).
41. Obayashi, E. *et al.* The structural basis for an essential subunit interaction in influenza virus RNA polymerase. *Nature* **454**, 1127–1131, doi:10.1038/nature07225 (2008).
42. Liu, J. *et al.* Structures of receptor complexes formed by hemagglutinins from the Asian Influenza pandemic of 1957. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 17175–17180, doi:10.1073/pnas.0906849106 (2009).
43. Gerritz, S. W. *et al.* Inhibition of influenza virus replication via small molecules that induce the formation of higher-order nucleoprotein oligomers. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 15366–15371, doi:10.1073/pnas.1107906108 (2011).
44. Russell, R. J. *et al.* The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **443**, 45–49, doi:10.1038/nature05114 (2006).
45. Xu, X., Zhu, X., Dwek, R. A., Stevens, J. & Wilson, I. A. Structural characterization of the 1918 influenza virus H1N1 neuraminidase. *J. Virol.* **82**, 10493–10501, doi:10.1128/JVI.00959-08 (2008).
46. Arzt, S. *et al.* Combined results from solution studies on intact influenza virus M1 protein and from a new crystal form of its N-terminal domain show that M1 is an elongated monomer. *Virology* **279**, 439–446, doi:10.1006/viro.2000.0727 (2001).
47. Hu, J. *et al.* Backbone structure of the amantadine-blocked trans-membrane domain M2 proton channel from Influenza A virus. *Biophys. J.* **92**, 4335–4343, doi:10.1529/biophysj.106.090183 (2007).
48. Hale, B. G. *et al.* Structural insights into phosphoinositide 3-kinase activation by the influenza A virus NS1 protein. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1954–1959, doi:10.1073/pnas.0910715107 (2010).
49. Hale, B. G., Barclay, W. S., Randall, R. E. & Russell, R. J. Structure of an avian influenza A virus NS1 protein effector domain. *Virology* **378**, 1–5, doi:10.1016/j.virol.2008.05.026 (2008).
50. Dias, A. *et al.* The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature* **458**, 914–918, doi:10.1038/nature07745 (2009).
51. Igarashi, M. *et al.* Predicting the antigenic structure of the pandemic (H1N1) 2009 influenza virus hemagglutinin. *PLoS One* **5**, e8553, doi:10.1371/journal.pone.0008553 (2010).
52. Yin, C. *et al.* Conserved surface features form the double-stranded RNA binding site of non-structural protein 1 (NS1) from influenza A and B viruses. *J. Biol. Chem.* **282**, 20584–20592, doi:10.1074/jbc.M611619200 (2007).
53. Meng, J., Zhang, Z., Zheng, Z., Liu, Y. & Wang, H. Methionine-101 from one strain of H5N1 NS1 protein determines its IFN-antagonizing ability and subcellular distribution pattern. *Sci. China. Life Sci.* **55**, 933–939, doi:10.1007/s11427-012-4393-9 (2012).
54. Dankar, S. K. *et al.* Influenza A/Hong Kong/156/1997(H5N1) virus NS1 gene mutations F103L and M106I both increase IFN antagonism, virulence and cytoplasmic localization but differ in binding to RIG-I and CPSF30. *Viol. J.* **10**, 243, doi:10.1186/1743-422X-10-243 (2013).

Acknowledgments

This work was supported by Contract No. HHSN266200700005C from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, and by ALSAC. We thank Dave Stallknecht, Scott Krauss, Dennis Senne, Tom DeLiberto and Seth Swafford for providing viruses; Angela McArthur for editorial assistance; Julie Groff for assistance with the graphic arts; and James Knowles for excellent administrative assistance.

Author contributions

Z.A.K. and R.G.W. conceived the study. Z.A.K. designed and performed the experiments. Z.A.K., Y.F., R.H., G.W. and J.O. analyzed the data. G.W., J.Z., R.W. and R.G.W. provided a supervisory role. All authors wrote and edited the manuscript.

Additional information

Supplementary information accompanies this paper at <http://www.nature.com/scientificreports>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Koçer, Z.A. *et al.* Survival analysis of infected mice reveals pathogenic variations in the genome of avian H1N1 viruses. *Sci. Rep.* **4**, 7455; DOI:10.1038/srep07455 (2014).



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>