**Parallel Evolution and the Emergence of Highly Pathogenic Avian Influenza A Viruses**

Marina Escalera-Zamudio[1], Bernardo Gutiérrez[1], Julien Thézé[1], Jeremy Russell Keown[2], Loic Carrique[2], Thomas A. Bowden[2] and Oliver G. Pybus[1]

1. Department of Zoology, Oxford University, Parks Rd Oxford, OX1 3PS, UK
2. Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Oxford, OX3 7BN, UK

## ABSTRACT

Surveillance of avian influenza viruses is crucial for the characterization of viral diversity in bird populations. Although virulent phenotypes are complex traits, several molecular determinants of pathogenicity in avian influenza viruses have been well characterised, particularly the polybasic proteolytic cleavage site within the Hemagglutinin protein, which allows a systemic spread of the infection in avian hosts. We hypothesise that the parallel evolution of highly pathogenic lineages, from low-pathogenic ancestors, may have been facilitated by permissive or compensatory secondary mutations occurring anywhere in the viral genome. We used a comparative phylogenetic and structural approach to detect mutations across the whole genome of avian influenza viruses of the H7NX subtype, that are shared by multiple, independent highly pathogenic lineages. Parallel mutations occur frequently among highly pathogenic outbreaks. Mutations strongly associated with the HP phenotype may function as an early detection system for transitional virulence stages.

## INTRODUCTION

Avian Influenza A viruses occur naturally among bird populations worldwide (Webster, et al. 1992). While most of these viruses cause only a mild to asymptomatic infection, highly pathogenic (HP) avian influenza A viruses (restricted to avian influenza viruses of the H5 and H7 subtype) can have an extremely high mortality rates of up to 100% (Webster, et al. 2006). Both low-pathogenic and HP viruses may cause severe disease in humans, as has been

documented during sporadic outbreaks resulting from avian-to-human and even from human-to-human transmissions (Buxton Bridges, et al. 2000) (Du Ry van Beest Holle, et al. 2005). As a recent example, the outbreak of H7N9 viruses in early 2018 caused over 1567 infections and 615 deaths among people in China (Lam and Pybus 2018). Low virulence lineages of avian influenza viruses can mutate to highly virulent strains, resulting geographically and temporally independent HP outbreaks that have generated huge economic losses in the poultry industry, and pose an increased risk for avian-to-human spread (Dhingra, et al. 2018) (Short, et al. 2015). Thus, understanding the mechanisms driving the evolution of virulence in avian influenza viruses is critical. Although virulent phenotypes are complex traits (Hatta, et al. 2001) (Jackson, et al. 2008) (Conenello, et al. 2007) (Schmolke, et al. 2011) (Qi, et al. 2014), one of the best characterized molecular markers of pathogenicity for avian influenza A viruses is a polybasic cleavage site (pCS) within the hemagglutinin (HA) protein, rendering the viral protein a target for broadly expressed host proteases that facilitate a systemic spread (Horimoto and Kawaoka 1994) (Abdelwhab, et al. 2016) (Hulse, et al. 2004).

Parallel molecular evolution describes repeated evolutionary changes leading to the same phenotype in unrelated populations, resulting from the adaptation by natural selection to comparable ecological niches (Stern 2013). The chances for adaptive evolution to occur increase as selective pressures and population sizes grow, such as during viral outbreaks (Stern 2013) (Pond, et al. 2012). Comparably, a combination of high rates of mutation and replication and a constrained genome size intrinsic to RNA viruses, coupled with consistently variable environments, results in fewer mutational pathways that can lead to a given phenotype, thus increasing the chances for parallel evolution to occur (Pond, et al. 2012). We hypothesize that the parallel evolution of HP lineages from low-pathogenic ancestors is characterized by the repeated emergence of permissive or compensatory secondary mutations occurring anywhere in the viral genome. Such mutations might arise prior to or immediately after the evolution of a polybasic cleavage site, in order to compensate detrimental effects on fitness, thus promoting evolvability of the HP phenotype. An analogous

process has been shown for the evolution of oseltamivir resistance in human H1N1 viruses (Bloom, et al. 2010). We developed a comparative approach integrating phylogenetics, molecular selection analysis and structural biology, to identify genome-wide mutations that exhibit parallel evolution in independent lineages, and we applied this approach to HP H7NX avian influenza A viruses. The advantage of such an approach compared to the frequently used contrasting presence/absence of molecular traits within individual HP outbreaks is that including a larger number of compared lineages should increase both the accuracy of detection of positive selection (Anisimova, et al. 2002) and helps avoids Type I errors in finding true genotype-to-phenotype links (Bhattacharya, et al. 2007).

We identified fifteen mutations within the PB2, PB1, NP, HA and NS1 gene segments and a deletion within NA gene (N1 and N3 subtypes), that evolved and emerged parallelly in distinct HP avian Influenza A lineages. Ranking these mutations according to their strong association to the HP phenotype generated a mutational panel that may function as an early detection system for transitional stages between low pathogenic to highly pathogenic phenotypes. Few of the mutations detected were inferred to be evolving under positive selection using standard techniques. Most involved conservative and fairly stabilising amino acid changes within effector domains in the solvent exposed surface of the proteins, but with no direct interactions with other residues or ligands. These results suggest that many of the parallel mutations detected may be selectively neutral or nearly neutral in the context of influenza virus genomes that lack a polybasic cleavage site conferring high virulence. It is therefore likely that these mutations fluctuate in frequency, through time and among locations, either through strong genetic drift or because they are linked to positively selected mutations elsewhere in the genome, such as the pCS.

## METHODS

### Large-scale phylogenetic analysis

All available nucleotide sequences for the H7 subtype viruses were downloaded from the Influenza Virus Resource (https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-

select.cgi?go=database) according to the following criteria: a minimum length of 1500 bases, removing identical sequences, including all geographical regions and all viral genomes that do not contain all segments. Sequences with more than 10% ambiguities were removed. The total number of sequences used for each genome segment dataset were the following: PB2 n= 1856, PB1 n= 1787, NP n= 1480, HA n= 2217 and NS1 n= 1231. Datasets was edited into the correct reading frame and aligned using MAFFT (Katoh, et al. 2002). Initial large-scale phylogenetic trees were estimated from nucleotide sequences using RAxML under the GTR+G model, and bootstrapping with 100 replicates (Stamatakis 2014). All HP clusters were flagged within the large-scale topologies according to the presence of more than one basic residue within the consensus cleavage site in the HA protein sequence, and by confirming their HP status with case reports in the literature (Dhingra, et al. 2018).

**Subsampling and cluster selection**

In order to carry out further analyses in a computationally feasible way, we subsampled the large-scale datasets. Selection of HP clusters to generate reduced phylogenies was done in a phylogenetically-informed manner, based on the HA large-scale tree. Genetically-distinct HP lineages were selected based on their segregation into well supported clusters (bootstrap values >80), the presence of ≥ 2 sequences, and on the availability of immediately ancestral sequences. Additionally, a subset of ≥ 150-300 sequences from the phylogeny backbone and a group of more distal ancestral sequences (for rooting purposes) were randomly selected. Equine and other non-avian host HP H7NX viruses were excluded, as it has been shown that these have host-specific independent rates of molecular evolution and could be a confounding factor our phylogenetic analyses (Worobey, et al. 2014). Datasets were aligned as described and were screened for recombination using the RDP, GENECOV and BOOTSCAN programs available in RDP3 package under default settings (Martin, et al. 2010). No evidence for recombination was detected by three or more methods under a P-value of < 0.001 (Bonferroni corrected). Final trees were estimated for each segment using RAxML under the GTR+G model and bootstrapping with 100 replicates (Stamatakis 2014).

The HA alignment was edited to remove the cleavage site region prior to phylogeny estimation. All nine genetically distinct HP lineages selected from the HA tree were supported by the clustering pattern observed in the other gene trees. The total number of sequences used for the reduced alignments were the following: PB2 n= 208, PB1 n=208, PA n=197, NP n=205, HA n=322, M1 n=159 and NS1 n=164. For the NA subtypes (total n=211): N1 n=45, N2 n=10, N3 n=84, N4 n=4, N6 n=2, N7 n=33, N8 n=4 and N9 n=28. The varying number of sequences used for each dataset and missing clusters within some of the gene trees is due to the lack of complete genomes for all strains. Accession numbers are available in the Supplementary File 1.

**Detecting parallel mutations and linking them to the HP phenotype**

Variable sites that were present in ≥1% of the sequences in the reduced alignments were selected to analyse their evolutionary history through reconstruction of ancestral states under the Maximum Likelihood framework, implemented by the "ace" function of the "ape" R package. The resulting phylogenetic trait reconstructions were visually inspected to determine any evolutionary patterns denoting parallel evolution. This was defined as amino acid changes shared by at least two HP clusters and which were present in at least 80% of the HP sequences, but not present the rest of the tree. In the case of the N segment, mutations within a single N subtype or within homologous sites among different N subtypes were considered. Global raw counts for the mutations of interest were obtained from the large-scale alignments using the WebLogo 3 online tool (http://weblogo.threeplusone.com/create.cgi). A $\chi^2$ test was used to determine the degree of association between each amino acid state and the HP phenotype using a 2x2 contingency table of the corresponding counts. The $\chi^2$ values were used to provide a ranking for the mutations of interest. Finally, for comparing global phylogenetic distribution patterns, sequences with the mutations of interest were extracted from the global alignments and back-traced to the original flagged trees. A non-random phylogenetic distribution is

considered when the sequences linked to a given mutation cluster within determined HP lineages, rather than being dispersed randomly throughout the tree.

**Analysis for molecular positive selection**

For the gene segments where parallel mutations were previously detected (PB2, PB1, HA, NP and NS1), global ω (dN/dS) values were estimated using CODEML in PAML v4 under the M1a/M2a site model (Yang 2007). To determine if parallel mutations associated with specific HP lineages had been a target of positive selection, individual tests were run using CODEML under the branch-site model A (BSA) (Yang 2007). HP clusters were individually constrained as foreground branches and only sites falling within the 2a categories (positive selection in selected branches and purifying selection for the rest of the tree) scored under BEB and NEB with a p-value ≥ 95% were considered (Yang 2007). In parallel, gene-wide (BUSTED; Branch-site Unrestricted Statistical Test for Episodic Diversification), site-specific (SLAC: Single-Likelihood Ancestor Counting and FEL; Fixed Effects Likelihood) and branch-site models (MEME; Mixed Effects Model of Evolution and BSREL; branch-site random effects likelihood) (Kosakovsky Pond and Frost 2005) (Murrell, et al. 2012) (Murrell, et al. 2015) (Smith, et al. 2015) implemented in the HyPhy package (Pond, et al. 2005) were used to test for pervasive and episodic selection acting a site, branch and branch-site level (Delport, et al. 2010). For all models, significance was assessed with LRTs against the null hypothesis using a $\chi^2$ distribution at a threshold of p-value ≤ 0.05, after correcting for multiple testing.

**Reconstruction of ancestral states for positively selected mutations**

For the genome segments were parallel mutations evolving under positive selection were detected (PB1, HA and NS1), we ran reconstructions of ancestral states for the sites of interest using BEAST v 1.8 (Drummond, et al. 2012). Trees were estimated using tip-dated datasets under the SRD06 nucleotide substitution model, a strict clock and a Bayesian Skyline Coalescent tree model. For each alignment, partitions were generated corresponding

to parallel mutations identified and were run under an asymmetric model, reconstructing ancestral states for all nodes. MCMC chains were run for $100x10^6$ or until convergence was reached for parameters of interest. Maximum clade credibility (MCC) trees were summarized from the posterior distribution using TreeAnnotator removing 10% of burn-in (Drummond, et al. 2012).

**Inferring the effects of mutations on protein stability and function**

To determine the location and hence functional relevance of the identified mutations, protein 3D structures were modelled from existing structural data available in the PDB (https://pdb101.rcsb.org/), using representative HP sequences from our datasets as a query. The high-resolution model from bat Influenza A/H17N10 (PDB ID:4WSB) was used as a template for the polymerase complex, as this is the only structural data available to date. The Influenza A/H5N1 nucleoprotein (PDB ID:2Q06) was used for the NS1 protein, and the crystal structure of a H7 influenza A virus hemagglutinin (PDB ID:3M5G) was used for the HA protein. For the HA and NS1 proteins, the I-TASSER server (Yang, et al. 2015) was used to generate models with the mutations of interest. These manually compared to those from the PDB for further validation. For the polymerase complex, mutations were introduced on the superimposed 3D protein structures using the PyMOL Molecular Graphics System v 2.1 (Schrödinger, LLC). Finally, to explore if the mutations of interest had an effect on global protein stability, the PIPS (Phylogenetic Inference of Protein Stability) program that uses a Bayesian approach with informative priors derived from physicochemical properties was employed (Bloom and Glassman 2009). PIPS estimates a value for free energy change (ddG) for each amino acid state in terms of the thermodynamic stability of a protein that unfolds and refolds, weighed under a phylogenetic background. The program was run for the PB2, PB1, HA and NS1 reduced datasets and ddG values were extracted for the sites of interest. Negative values are interpreted as fairly stabilizing/neutral mutations, whilst values of ≤-5 are indicative of highly stabilizing mutations (Bloom and Glassman 2009).

## RESULTS

### Evidence for parallel evolution in independent HP outbreaks

Large-scale phylogenetic analysis confirmed the divergence of H7NX viruses in two major geographically defined clades (the Eurasian and North American lineages) (Xu, et al. 2016) (Banks, et al. 2001). Historical HP outbreaks corresponding to independent geographically and temporally defined lineages displayed for most of the cases a unique pCS sequence (Figure 1A). Low-pathogenic H7NX viruses showed a consensus sequence of -PEXPKGRGLF- within the cleavage site, whereas a minimal consensus -PEXPKXnK/RnRGLF- was identified for the HP viruses (Bosch, et al. 1981) (Abdelwhab, et al. 2016). Nine different genetically distinct HP clusters (C1-C9) were selected from the inferred HA phylogeny (Figure 1B). These clusters also represented monophyletic clades in the rest of the genome segments phylogenies (Supplementary Figure 1 and 2). Clustering of HP lineages was generally sustained within all phylogenies, with the exception of C1, which segregated into C1.1 and C1.2 in the PB2, PB1, PA, M and NS1 trees. Clusters missing in some of the phylogenies is due to the lack of corresponding sequences for all viral genes. Although reassortment is frequent among avian Influenza A viruses (Lu, et al. 2014), the consistent clustering of HP lineages in all trees provided little evidence that reassortment occurs within the HP lineages.

For each genome segment, we detected between 21 and 125 variable sites present in at least 1% of the sequences subsampled, with the HA segment being the most variable. Reconstruction of ancestral amino acid states revealed fifteen candidate parallel mutations present in at least two HP clusters and the immediate ancestral nodes, but not fixed in the rest of the tree (Supplementary File 2). For PB2, mutations R355K and V80I shared by C5 and C6 were detected. For PB1, mutation S152L shared by C6 and C9 and D154 in C2 and C6 were identified. For HA (in H7 numbering), mutation S136N shared by C1 and C8, A143T in C2, C5, C6, C7 and C8, M/L274I in C4 and C8, D384N in C6 and C8, and V438I in C3 and C8 were identified. Similarly, mutation S377N shared by C1, C2 and C8 was found in NP. Finally, mutation T80S shared by C1.1 and C1.2, N127T in C1.1 and C6, D139N and V180I

in C1 and C2, and D209N in C1.2, C6 and C8 were detected in NS1. We also identified a 22 amino acid deletion within stalk region of the NA (N1 and N3 subtypes, sites 54-75 in N3 numbering) associated with C2 and C7 (Table 1).

**Structural analysis of parallel mutations**

Of the parallel mutations detected, non-conservative amino acid changes or mutations within functionally relevant regions of the viral proteins include: R355K in PB2, S152L and G154D/S in PB1, S136N and A143T in HA, and N127T, D139N, V180I and D209N in NS1 (Figure 2). Mutation R355K in PB2 is a conservative amino acid change within a solvent exposed loop of the electrostatic surface of cap binding domain. This site is in direct contact with first phosphates after the cap (Guilligay, et al. 2008). Mutations S152L and G154D/S in PB1 are non-conservative amino acid changes within the core domain of the protein, but with no known important function. Mutation S136N is associated with additional glycosylation known to facilitate evasion of Ab-mediated immunity (She, et al. 2017), whilst A143T in HA is a non-conservative amino acid change located in a solvent-accessible region within the 130-loop proximal to the receptor surface. This site located within the antigenic pocket A, but has no direct interaction with glycan ligands (Yang, et al. 2010). Mutation N127T in NS1 is a conservative change involved in protein kinase R activation (Mok, et al. 2012). Similarly, D139N is a conservative change located within the nuclear export signal (Li, et al. 1998). This site also forms part a loop-forming Interface B (residues 139-143) that directly interacts with the central four-helical bundle of the antiviral protein TRIM25-CC (Koliopoulos, et al. 2018). Mutation V180I is located within the cleavage and polyadenylation specificity factor (CPSF) involved in the processing of pre-mRNAs (Mandel, et al. 2006). This site is likely to interact with CPSF30, as inferred from the high-resolution atomic structure of the human polyadenylation complex (Sun, et al. 2018). Mutation N209 is located within a disordered tail interacting with importin alpha, which promotes the entry of viral proteins to the nucleus and is essential for virus replication (Melen, et al. 2007). Finally, the 22 amino acid deletion found in NA is located within the stalk region of the protein. Although such a deletion does not have

a direct functional effect, the conformation of the neuraminidase enzymatic pocket can differ depending on stalk length, with possible repercussions on its binding properties (Durrant, et al. 2016). Similarly, stalk deletions in NA are associated with increased glycosylation patterns in HA, shown to reduce receptor binding affinity (Li, et al. 2011). Most of the parallel mutations detected represent conservative amino acid changes within core regions of the proteins and with no inferred direct residue-residue or ligand-residue interactions. These mutations also have mostly negative ddG values, rendering them as fairly stabilizing/neutral changes (Figure 3). However, we cannot rule out that the parallel mutations detected are of biological relevance, as epistatic interactions with other viral proteins and cellular factors cannot be inferred from these structural analyses.

**Molecular selection analysis of parallel mutations**

Branch, site and branch-site model testing using independent methods provided evidence for episodic and pervasive selection acting upon some of the parallel mutations associated with HP clusters, and upon specific HP lineages (C4 and C7) (Table 2, Supplementary File 3 and 4). Parallel mutations that were identified as evolving under positive selection include S152L in PB1, A143T in HA and V180I and D209N in NS1 (Figure 3). Positively selected sites correlate to functionally relevant regions of the viral proteins, strengthening the relevance of their association with specific HP lineages. No evidence for gene-wide pervasive positive selection was detected, suggesting that the proportion of positively selected sites is very small compared to that of sites evolving under nearly neutral or purifying selection (Suzuki 2006), consequently returning global ω (dN/dS ratio) values of < 1. Whilst positive selection plays a crucial role in fixing beneficial and compensatory mutations, permissive mutations are likely fixed through linkage and soft sweeps from the standing genetic variation (Frost, et al. 2018). Thus, we show evidence that most parallel mutations, if they adaptive, represent permissive changes that can occur anywhere throughout the viral genome and are fixed through linkage and genetic sweeps before an outbreak. On the other hand, the role of selection acting upon individual HP lineages is less clear, as there is no evidence that the HP

phenotype is advantageous and therefore favoured compared to the low-pathogenic variants (Saenz, et al. 2012). For example, mathematical modelling has been previously used to hypothesise that the population dynamics of HP strains may be determined by their linkage to one or more specific antigenic variants, whose frequencies are driven by the dynamics of host immunity, rather than by mutability (Wikramaratna, et al. 2014).

**Strong link between parallel mutations and the HP phenotype**

In order to detect genotype-to-phenotype links, we determined the degree of association between each amino acid state of interest and the HP phenotype, and further explored the global phylogenetic distribution of the sequences linked to parallel mutations within the large-scale trees. A strong association between most of the detected mutations with the HP phenotype was supported by both $\chi^2$ values and by an observed non-random phylogenetic distribution denoted by clustering linked to the HP phenotype (Table 1). We acknowledge that the $\chi^2$ test is not fully statistically valid, as we cannot dismiss count interdependency due to possible phylogenetic correlations (Bhattacharya, et al. 2007); thus, $\chi^2$ values could be overestimated. Correcting for this error implies developing complex models that are not readily available, and goes beyond the scope of this study (Ansari and Didelot 2016). However, it would be important to implement such models within our pipeline in the future to further correct such possible biases. To strengthen our ranking, we further added information on the biological properties of the mutations of interest, taking into account if these are in functionally relevant regions of the viral proteins, if they are evolving under positive selection associated to HP lineages, if they involve non-conservative amino acid changes or if they comprise stabilizing/neutral changes. Considering all variables, the strongest genotype-to-phenotype association was found for mutations A143T and D384N in HA, G154D/S and S152L in PB1, R355K and V480I in PB2, 377N in NP and D139N, V180I and D209N in NS1 (Table 1). No genotype-to-phenotype association was detected for mutations S136N and V438I in HA and T80S in NS1. A strong association to HP was observed for the NA deletion

in the N1 subtype, but not for the N3 subtype. Despite such deletion being previously linked the C2 viruses, other deletion patterns have also been repeatedly observed in six out of nine NA subtypes, including low-pathogenic H6NX avian Influenza A viruses (Li, et al. 2011). It has been suggested that NA stalk deletions in avian influenza A viruses are strongly linked to the adaptation to domestic poultry (Munier, et al. 2010), as supported by in vitro observations that the stalk truncation enhances replication and virulence of influenza virus in eggs and chickens (Sorrell, et al. 2010). Therefore, the direct association of NA stalk deletions with the HP phenotype remains unclear.

## DISCUSSION

Previous studies had pointed out parallel mutations (within the HA gene) in independent H7NX avian influenza A viral lineages, but had not associated these with the HP phenotype. Similarly, other studies had observed mutations linked to case-specific highly pathogenic outbreaks, but had not identified such mutations to be evolving parallely (Lebarbenchon and Stallknecht 2011), (Monne, et al. 2014), (Fusaro, et al. 2015). For example, mutations A143T and S136N were characterized in HP H7N9 viruses circulating in China during 2017-2018 (corresponding to the C1 viruses here) (Lu., et al. 2018). Similarly, mutation A143T was also detected in HP viruses form Pakistan (corresponding to C6), and in HP viruses from the Italy 1999-2001 outbreak (corresponding to C2) (Lebarbenchon and Stallknecht 2011), (Monne, et al. 2014). Mutations G154D/S in PB1, S377N in NP and D139N in NS1 were also previously associated with the emergence of HP viruses from low pathogenic progenitors during the Italy 1999-2001 outbreak (Monne, et al. 2014), together with the 22 amino acid deletion in the NA gene (Banks, et al. 2001). Similarly, mutation R355K in PB2 was detected in two epidemiologically linked outbreaks of low and highly pathogenic H7N7 viruses in Germany during 2105 (Dietze, et al. 2018). Through our methodological pipeline, we have encompassed independent findings described in such studies, proving the efficacy of using a comparative phylogenetic approach to systematically provide evidence for parallel evolution within HP independent lineages. Moreover, some of these mutations seem to be shared with

HP H5NX viruses (unpublished data), supporting the observation that a constrained genome size and similar selective pressures within comparably changing environments, result in restricted genomic pathways leading to the emergence of the HP phenotype. As parallelisms are common in avian influenza A viruses, the evolution of specific phenotypes can be forecasted, to some extent, by understanding repeating mutational patterns.

It is known that mutations with a functional impact almost invariably carry a fitness cost, but are frequently compensated by subsequent fitness restoring or permissive mutations (Frost, et al. 2018). However, in the absence of experimental data it is not possible to evaluate whether if the parallel mutations detected here are selectively neutral or advantageous in the context of the HP phenotype. Since most of these are positively associated with the HP phenotype, we can assume that they are not deleterious in the presence of a polybasic cleavage. Nonetheless, further experimental validation would be necessary to elucidate what is the role of these mutations in the emergence of the virulent phenotype. In conclusion, the mutational panel provided here may function as an early detection system for transitional stages between low pathogenic to highly pathogenic phenotypes. Circulating viral strains in domestic and wild bird populations that do not have a polybasic cleavage site yet, but show all or some of the amino acid changes ranked according to our mutational panel, should remain under close vigilance as potential HP mutants.

**TABLE 1. Parallel mutations ranked by their association with the HP phenotype**

| Rank | Mutation | Segment | Shared by clusters | GF[1] | GF-HP[2] | GF-LP[3] | Chi-Squared | P-value | GPD-HP[4] | PSS[6] |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | A143T | HA | C2, C5, C6, C7 and C8 | A 84%, T 7% | A 18%, T 55% | A 90%, T 2% | 1030.1 | 0.00 | Yes | Yes |
| 2 | G154D | PB1 | C2 and C6 | G95%, D4% | G44%, D52% | G99%, D 0.43% | 582.1 | 0.00 | Yes | No |
| 3 | D384N | HA | C6 and C8 | D 96%, N 4% | D 74%, N 26% | D 98%, N 2% | 296.0 | 0.00 | Yes | No |
| 4 | R355K | PB2 | C5 and C6 | R 98%, K 2% | R 82%, K 11% | R 99%, K 0.3% | 107.6 | 0.00 | Yes | No |
| 5 | V480I | PB2 | C5 and C6 | V 98%, I 2% | V 86%, I 14% | V 99%, I 0.75% | 91.8 | 0.00 | Yes | No |
| 6 | S377N | NP | C1, C2 and C8 | S 70%, N 30% | S 39%, N 61% | S 79%, N 20% | 71.4 | 0.00 | No | No |
| 7 | D209N | NS1 | C1.1, C6 and C8 | D 71%, N 26% | D 32%, N 64% | D 74%, N 26% | 62.8 | 0.00 | Yes | Yes |
| 8 | D139N | NS1 | C1 and C2 | D 68%, N 29% | D 45%, N 49% | D 76%, N 20% | 35.9 | 0.00 | Yes | No |
| 9 | G154S | PB1 | C2 and C6 | G95%, S1% | G44%, S4% | G98%, S 0.69% | 28.1 | 0.00 | Yes | No |
| 10 | S152L | PB1 | C6 and C9 | S 98%, L 2% | S 92%, L 8% | S 99%, L 1% | 27.8 | 0.00 | Yes | Yes |
| 11 | V180I | NS1 | C1 and C2 | V 54%, I 44% | V 28%, I 59% | V 61%, I 39% | 21.3 | 0.00 | Yes | Yes |
| 12 | M/L274I | HA | C4 and C8 | M/L 95%, I 5% | M/L 90%, I 9% | M/L 95%, I 4% | 14.2 | 0.00 | No | No |
| 13 | N127T | NS1 | C1.1 and C7 | N 68%, T 8% | N 40%, T 10% | N 70%, T 5% | 10.3 | 0.00 | Yes | No |
| 14 | V438I | HA | C3 and C8 | V 96%, I 4% | V 93%, I 6% | V 96%, I 4% | 2.6 | 0.07 | No | No |
| 15 | S136N | HA | C1 and C8 | S 91%, N 8% | S 78%, N 9% | S 92%, N 7% | 1.5 | 0.15 | No | No |
| 16 | T80S | NS1 | C1.1 and C1.2 | T 58%, S 2% | T 37%, S 8% | T 63%, S 12% | 0.2 | 0.86 | No | No |

[1]GF: Global Frequencies for different amino acid states derived from the raw counts obtained from the total number of sequences available in NCBI database.
[2]GF-HP: Global Frequency for different amino acid states in the HP sequences
[3]GF-LP: Global Frequency for different amino acid states in the LP sequences
[4]GPD-HP: Non-randon Global Phylogenetic Distribution associated with HP lineages
[5]PSS: positively selected site

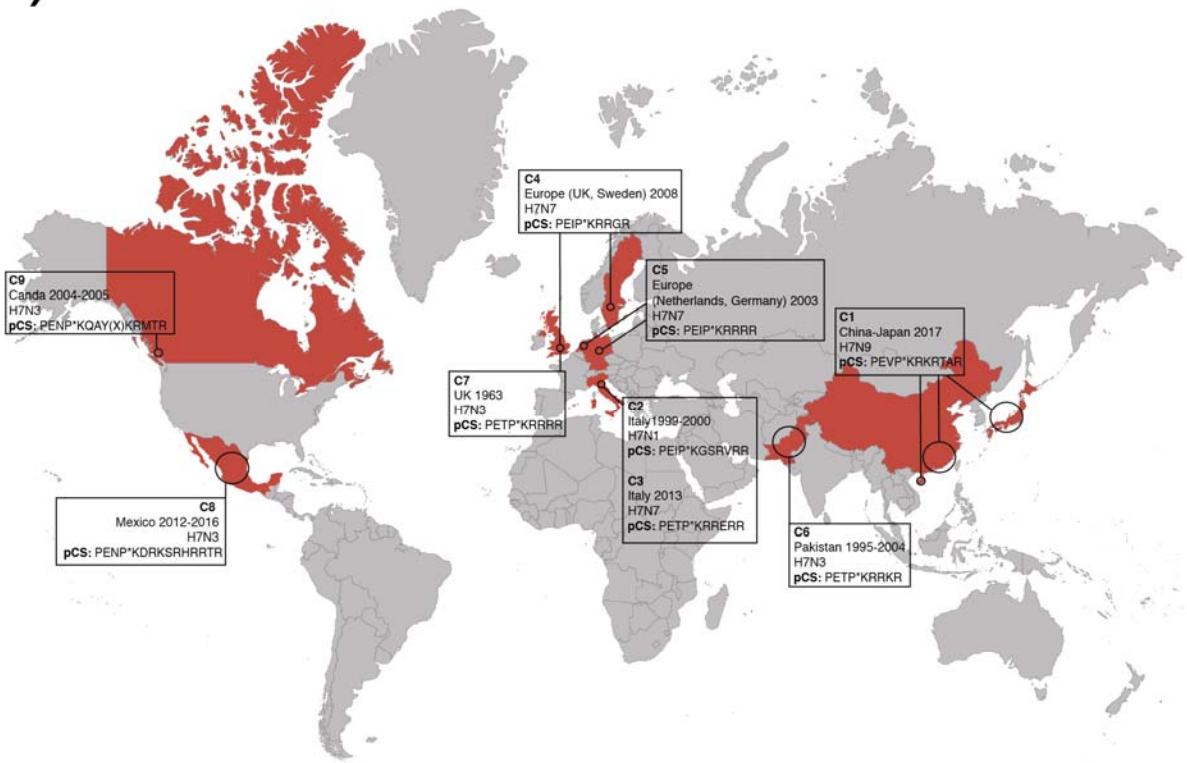
**TABLE 2. Positive selection acting upon paralllel mutations associated with HP lineages**

| Tests for gene-wide pervasive selection | | | | Site models for pervasive selection | | | | Site models for episodic selection | | Branch-site models for episodic selection | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Gene** [a] | **BUSTED ≤ 0.05** | **M1a/M2a ≤ 0.05** | **Mean *ω* *(dN/dS)*** [b] | **SLAC ≤ 0.05** | **SITE** [a] | **FEL ≤ 0.05** | **SITE** [a] | **MEME ≤ 0.05** | **SITE** [a] | **BSA ≤ 0.05** [c] | **CLUSTER** | **SITE** [a,d] | **bsREL ≤ 0.05** | **CLUSTER** | **SITE** |
| PB2 | NO | NO | 0.028 | NO | -- | YES | -- | YES | -- | NO | -- | **355** | NO | -- | -- |
| PB1 | NO | NO | 0.035 | NO | -- | YES | **152** | YES | -- | NO | C9 | **152** | NO | -- | -- |
| H7 | YES | NO | 0.085 | NO | -- | YES | **143** | YES | -- | YES | **C4** | -- | YES | **C4, C7** | -- |
| NP | NO | NO | 0.030 | NO | -- | NO | -- | YES | -- | NO | -- | -- | NO | -- | -- |
| N3 | NO | NO | 0.090 | NO | -- | YES | -- | NO | -- | NO | -- | -- | NO | -- | -- |
| NS | NO | NO | 0.164 | NO | -- | YES | **180, 209** | YES | **180, 209** | NO | C8 | **180** | NO | -- | -- |

[a] The genes/sites shown in this table correspond only to those where parallel mutations were detected and positively selected sites scored as parallel mutations by reconstruction of ancestral states.
[b] The mean ω (dN/dS) estimates were taken from the M1a/M2a test.
[c] Although the significance for the BSA test can be negative due to very small global proportion of PSS, sites can still be significantly scored under BEB/NEB
[d] PSS scored by both BEB/NEB with a P-value ≥ 95%

**Figure 1.**

A)


C4
Europe (UK, Sweden) 2008
H7N7
pCS: PEIP*KRRGR
C5
Europe
(Netherlands, Germany) 2003
H7N7
pCS: PEIP*KRRRR
C9
Canda 2004-2005
H7N3
pCS: PENP*KQAY(X)KRMTR
C1
China-Japan 2017
H7N9
pCS: PEVP*KRKRTAR
C7
UK 1963
H7N3
pCS: PETP*KRRRR
C2
Italy1999-2000
H7N1
pCS: PEIP*KGSRVRR
C3
Italy 2013
H7N7
pCS: PETP*KRRERR
C8
Mexico 2012-2016
H7N3
pCS: PENP*KDRKSRHRRTR
C6
Pakistan 1995-2004
H7N3
pCS: PETP*KRRKR

B)

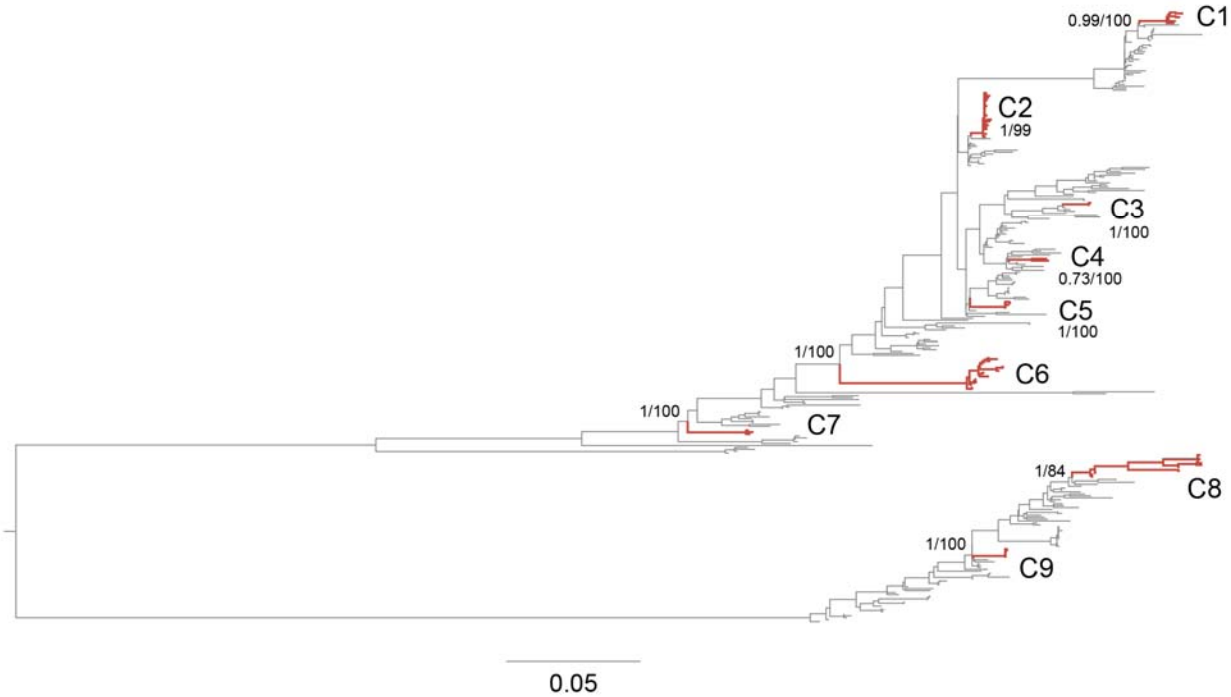

0.99/100
C1
C2
1/99
C3
1/100
C4
0.73/100
C5
1/100
1/100
C6
1/100
C7
1/84
C8
1/100
C9
0.05

**Figure 2.**

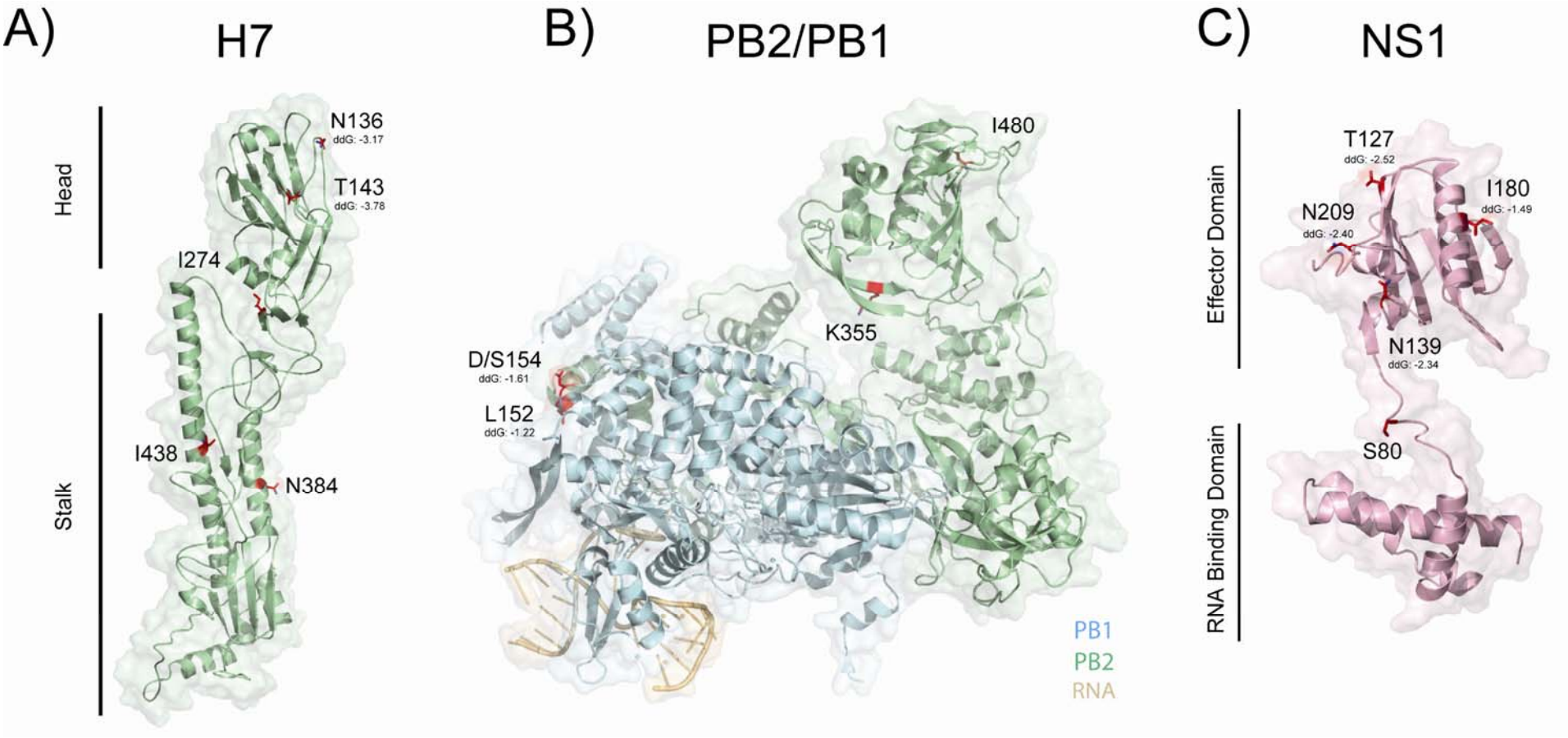
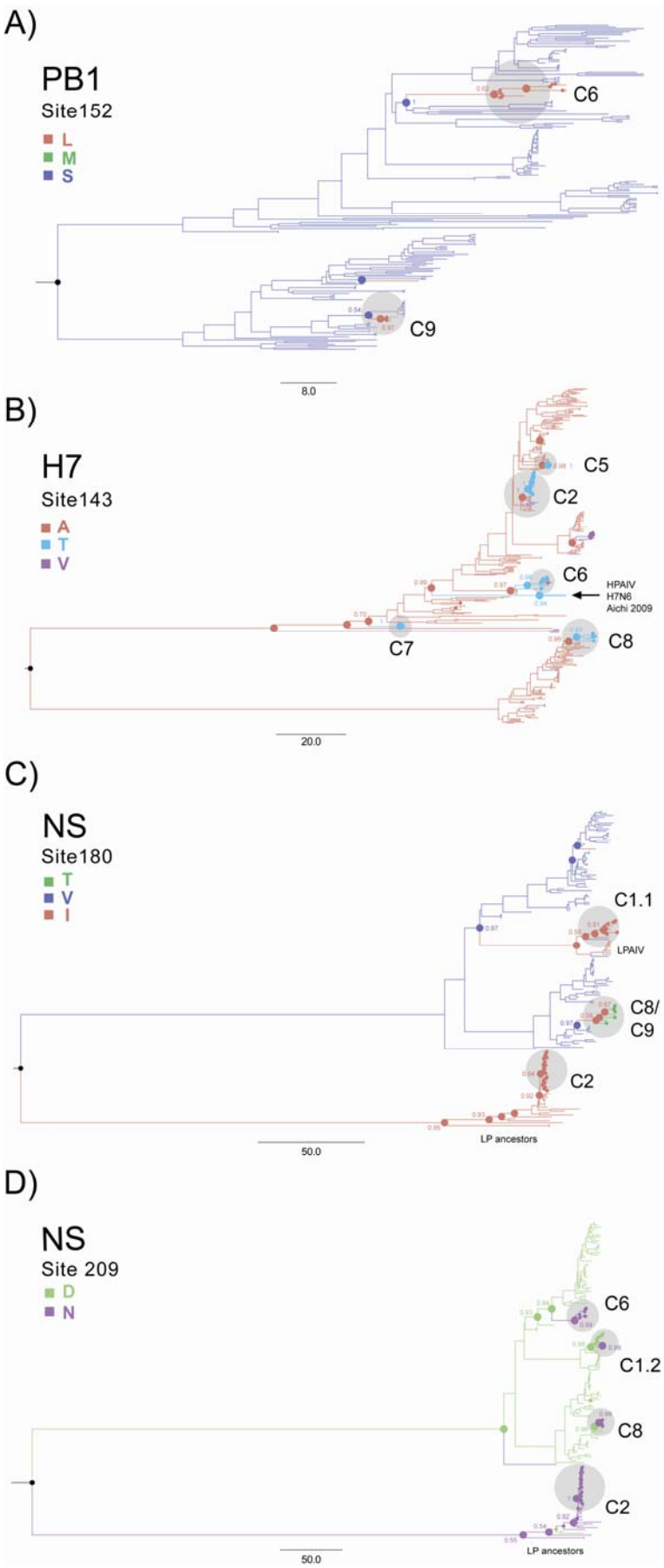
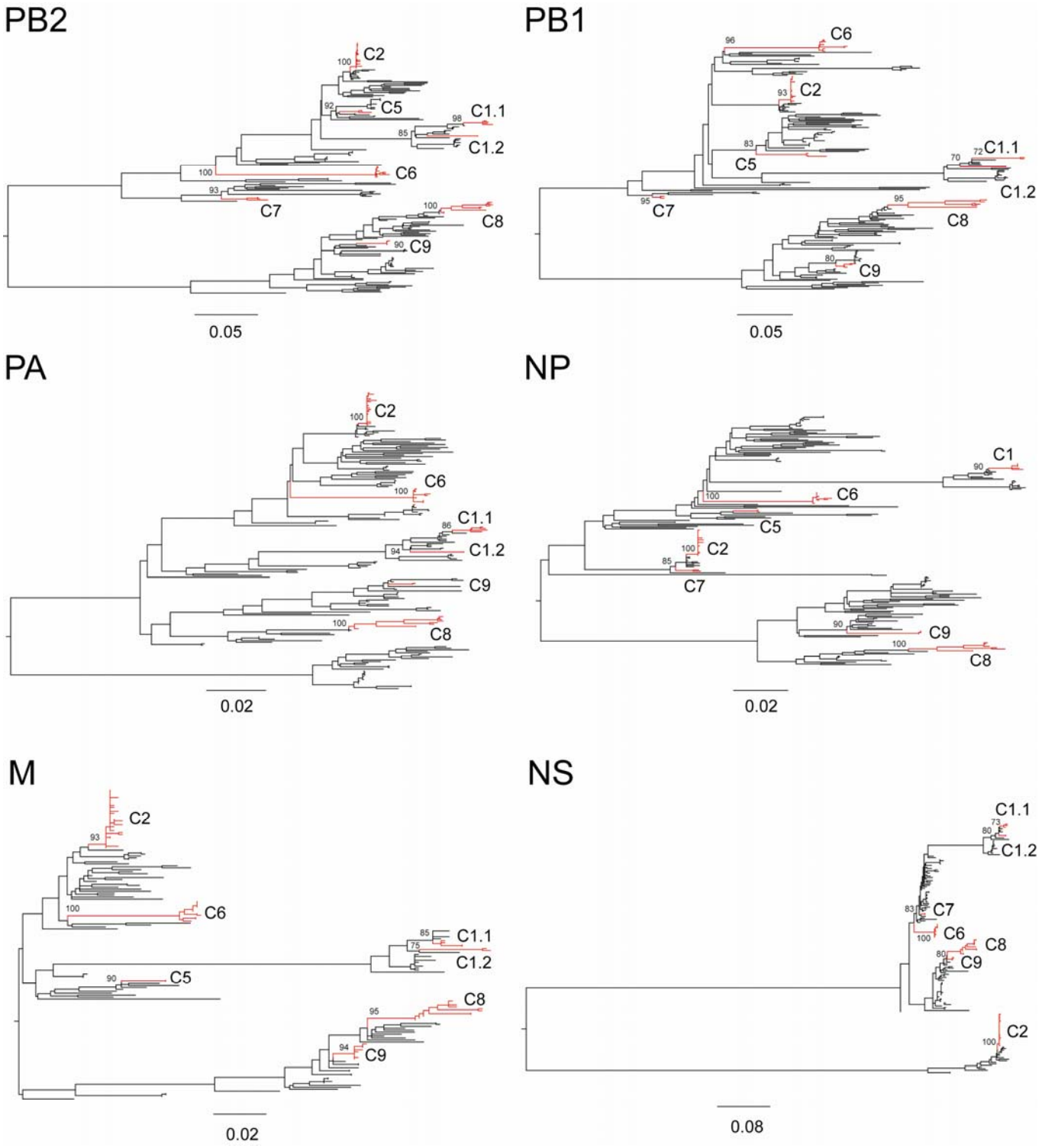

A)
H7
Head
Stalk
N136
ddG: -3.17
T143
ddG: -3.78
I274
I438
N384
B)
PB2/PB1
I480
K355
D/S154
ddG: -1.61
L152
ddG: -1.22
PB1
PB2
RNA
C)
NS1
Effector Domain
RNA Binding Domain
T127
ddG: -2.52
I180
ddG: -1.49
N209
ddG: -2.40
N139
ddG: -2.34
S80

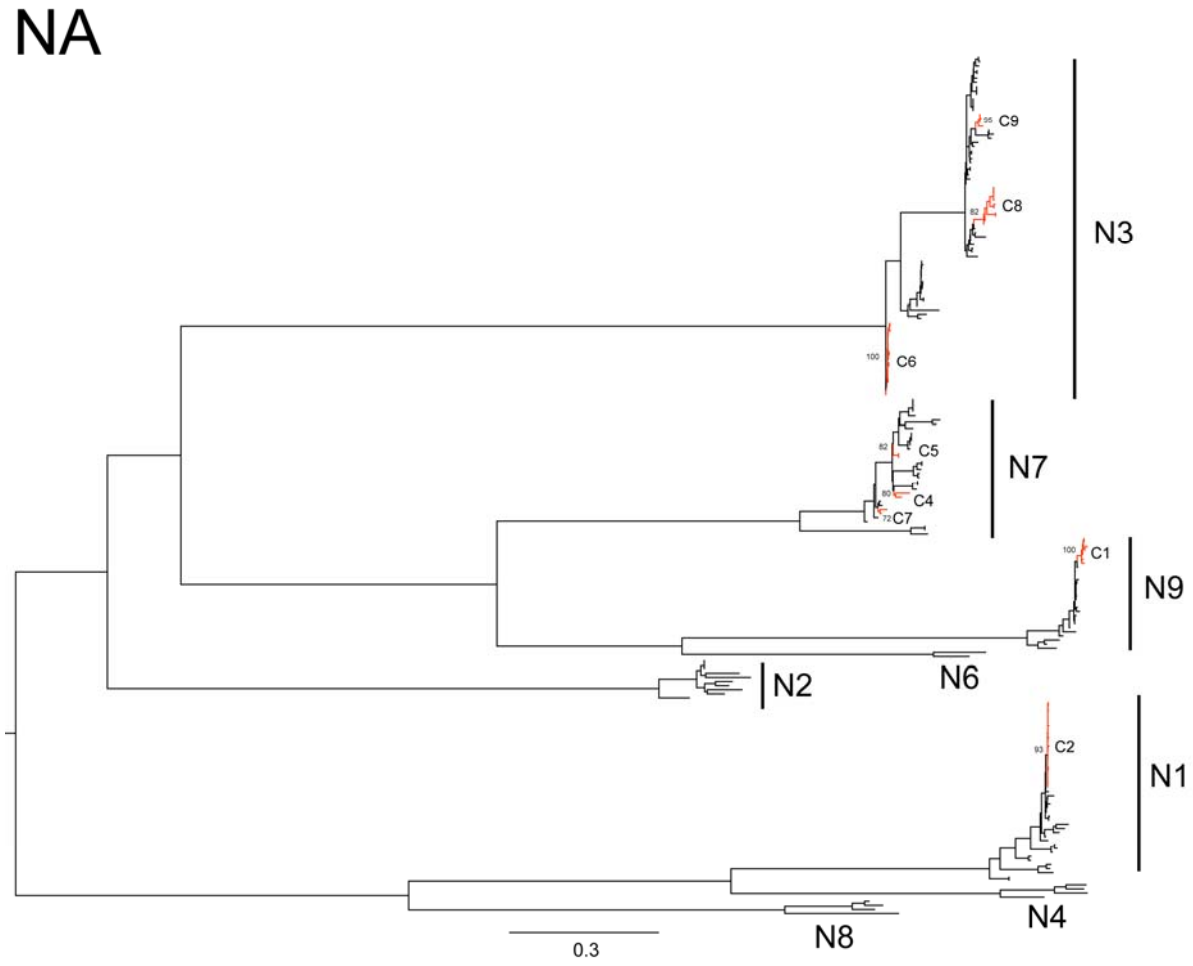
**Figure 3.**


A)
PB1
Site152
L
M
S
C6
C9
8.0
B)
H7
Site143
A
T
V
C5
C2
C6
HPAIV
H7N6
Aichi 2009
C7
C8
20.0
C)
NS
Site180
T
V
I
C1.1
LPAIV
C8/
C9
C2
LP ancestors
50.0
D)
NS
Site 209
D
N
C6
C1.2
C8
C2
LP ancestors
50.0

**Supplementary Figure 1.**



**Supplementary Figure 2.**



## FIGURE LEGENDS

**Figure 1. Historical and geographical occurrence of H7NX highly pathogenic (HP) clusters.**

A) Map representation of the geographical occurrence of historical HP outbreaks. The country and year of circulation, viral subtype and the sequence of the polybasic cleavage site (pCS) are indicated for each of the given cluster (C1-C9). B) Beast tree for the HA gene representing the nine genetically and phenotypically distinct clusters selected for this study (C1-C9, highlighted in red branches). Posterior probability and bootstrap values are indicated for the nodes of interest.

**Figure 2. 3D protein structure models of the H7, PB2/ PB1 complex and NS1 viral proteins.**

Structural models showing the parallel mutation identified (highlighted in red). Only negative ddG values are indicated. A) For HA, mutations S136N and A143T in HA are located within a solvent-accessible region close to the receptor binding surface. M274I is located within the receptor-binding subdomain, whilst D384N and V438I are located within membrane-fusion inducing stalk domain. B) For PB2, mutation R355K is within a solvent exposed loop of the electrostatic surface of cap binding domain, whilst mutation V480I is close to the nuclear localisation signal. For PB1, mutations S152L and G154D/S are within the core domain of the protein. C) For NS1, mutation N127T is within a solvent exposed region involved in protein kinase R activation. Similarly, D139N is located within the nuclear export signal and forms part a loop-forming Interface that interact with the antiviral protein TRIM25. Mutation V180I is located within the cleavage and polyadenylation specificity factor (CPSF) and thus is likely to interact with the cellular CPSF30. Finally, mutation N209 is located within a disordered tail interacting with host importins.

**Figure 3. Reconstruction of ancestral states for positively selected parallel mutations.**

BEAST trees denoting the reconstruction of ancestral amino acid states for positively selected parallel mutations. Branches within the trees are coloured according to the corresponding amino acid stated in the legend. Ancestral nodes preceding the emergence of an amino acid change associated to the emergence of HP lineages are represented with coloured circles. The probability of a given amino acid state occurring in specific nodes is indicated. Clusters of interest are highlighted with grey circles. A) S152L in PB2 is shared by C6 and C9. B) Mutation A143T in HA is shared by C2, C5, C6, C7 and C8 and also occurs in other HP sequences not considered for this study. C) Mutation V180I in NS1 is shared by C1.1, C2, C8 and C9. This change also occurs in the low pathogenic ancestors to C2. D) Mutation D209N in NS1 is shared by C1.2, C2 C6 and C8 and occurs in the low pathogenic ancestors to C2.

**Supplementary Figure 1. Trees for the PB2, PB1, PA, NP, M1 and NS1 genome segments**

ML trees for the PB2, PB1, PA, NP, M1 and NS1 viral genome segments showing the divergence of HP lineages in accordance to the clustering pattern observed in the HA tree. HP clusters (C1-C9) are indicated with red branches and bootstrap values are indicated for the nodes of interest. Distances are shown in nucleotide changes/site/year.

**Supplementary Figure 2. Tree NA subtype ML trees for the H7NX viruses**

ML trees for the NA viral genome segment (N1-N9 subtypes) showing the divergence of HP lineages in agreement to the clustering pattern observed in the HA. HP clusters (C1-C9) are indicated with red branches and bootstrap values are indicated for the nodes of interest. Distance is shown in nucleotide changes/site/year.

## SUPPLEMENTARY FILES

**Supplementary File 1. Accession numbers for reduced datasets**

H7_ACCESSIONS.xls

**Supplementary File 2. Reconstruction of ancestral states ML trees**

RAS_MLsites.pdf

**Supplementary File 3. Likelihood Ratio Tests for CODEML**

H7_LRT_CODEML.xls

**Supplementary File 4. Positively Selected Sites scored under BST/CODEML**

H7_PSS_CODEML_BST.xls

## REFERENCES


Abdelwhab EM, Veits J, Ulrich R, Kasbohm E, Teifke JP, Mettenleiter TC. 2016. Composition of the Hemagglutinin Polybasic Proteolytic Cleavage Motif Mediates Variable Virulence of H7N7 Avian Influenza Viruses. Sci Rep 6:39505.

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950-958.

Ansari MA, Didelot X. 2016. Bayesian Inference of the Evolution of a Phenotype Distribution on a Phylogenetic Tree. Genetics 204:89-98.

Banks J, Speidel ES, Moore E, Plowright L, Piccirillo A, Capua I, Cordioli P, Fioretti A, Alexander DJ. 2001. Changes in the haemagglutinin and the neuraminidase genes prior to the emergence of highly pathogenic H7N1 avian influenza viruses in Italy. Arch Virol 146:963-973.

Bhattacharya T, Daniels M, Heckerman D, Foley B, Frahm N, Kadie C, Carlson J, Yusim K, McMahon B, Gaschen B, et al. 2007. Founder effects in the assessment of HIV polymorphisms and HLA allele associations. Science 315:1583-1586.

Bloom JD, Glassman MJ. 2009. Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. PLoS Comput Biol 5:e1000349.

Bloom JD, Gong LI, Baltimore D. 2010. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. Science 328:1272-1275.

Bosch FX, Garten W, Klenk HD, Rott R. 1981. Proteolytic cleavage of influenza virus hemagglutinins: primary structure of the connecting peptide between HA1 and HA2 determines proteolytic cleavability and pathogenicity of Avian influenza viruses. Virology 113:725-735.

Buxton Bridges C, Katz JM, Seto WH, Chan PK, Tsang D, Ho W, Mak KH, Lim W, Tam JS, Clarke M, et al. 2000. Risk of influenza A (H5N1) infection among health care workers exposed to patients with influenza A (H5N1), Hong Kong. J Infect Dis 181:344-348.

Conenello GM, Zamarin D, Perrone LA, Tumpey T, Palese P. 2007. A single mutation in the PB1-F2 of H5N1 (HK/97) and 1918 influenza A viruses contributes to increased virulence. PLoS Pathog 3:1414-1421.

Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455-2457.

Dhingra MS, Artois J, Dellicour S, Lemey P, Dauphin G, Von Dobschuetz S, Van Boeckel TP, Castellan DM, Morzaria S, Gilbert M. 2018. Geographical and Historical Patterns in the Emergences of Novel Highly Pathogenic Avian Influenza (HPAI) H5 and H7 Viruses in Poultry. Front Vet Sci 5:84.

Dietze K, Graaf A, Homeier-Bachmann T, Grund C, Forth L, Pohlmann A, Jeske C, Wintermann M, Beer M, Conraths FJ, et al. 2018. From low to high pathogenicity-Characterization of H7N7 avian influenza viruses in two epidemiologically linked outbreaks. Transbound Emerg Dis.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969-1973.

Du Ry van Beest Holle M, Meijer A, Koopmans M, de Jager CM. 2005. Human-to-human transmission of avian influenza A/H7N7, The Netherlands, 2003. Euro Surveill 10:264-268.

Durrant JD, Bush RM, Amaro RE. 2016. Microsecond Molecular Dynamics Simulations of Influenza Neuraminidase Suggest a Mechanism for the Increased Virulence of Stalk-Deletion Mutants. J Phys Chem B 120:8590-8599.

Frost SDW, Magalis BR, Kosakovsky Pond SL. 2018. Neutral Theory and Rapidly Evolving Viral Pathogens. Mol Biol Evol 35:1348-1354.

Fusaro A, Tassoni L, Hughes J, Milani A, Salviato A, Schivo A, Murcia PR, Bonfanti L, Cattoli G, Monne I. 2015. Evolutionary trajectories of two distinct avian influenza epidemics: Parallelisms and divergences. Infect Genet Evol 34:457-466.

Guilligay D, Tarendeau F, Resa-Infante P, Coloma R, Crepin T, Sehr P, Lewis J, Ruigrok RW, Ortin J, Hart DJ, et al. 2008. The structural basis for cap binding by influenza virus polymerase subunit PB2. Nat Struct Mol Biol 15:500-506.

Hatta M, Gao P, Halfmann P, Kawaoka Y. 2001. Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. Science 293:1840-1842.

Horimoto T, Kawaoka Y. 1994. Reverse genetics provides direct evidence for a correlation of hemagglutinin cleavability and virulence of an avian influenza A virus. J Virol 68:3120-3128.

Hulse DJ, Webster RG, Russell RJ, Perez DR. 2004. Molecular determinants within the surface proteins involved in the pathogenicity of H5N1 influenza viruses in chickens. J Virol 78:9954-9964.

Jackson D, Hossain MJ, Hickman D, Perez DR, Lamb RA. 2008. 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.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059-3066.

Koliopoulos MG, Lethier M, van der Veen AG, Haubrich K, Hennig J, Kowalinski E, Stevens RV, Martin SR, Reis ESC, Cusack S, et al. 2018. Molecular mechanism of influenza A NS1-mediated TRIM25 recognition and inhibition. Nat Commun 9:1820.
Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208-1222.

Lam TT, Pybus OG. 2018. Genomic surveillance of avian-origin influenza A viruses causing human disease. Genome Med 10:50.

Lebarbenchon C, Stallknecht DE. 2011. Host shifts and molecular evolution of H7 avian influenza virus hemagglutinin. Virol J 8:328.

Li J, Zu Dohna H, Cardona CJ, Miller J, Carpenter TE. 2011. Emergence and genetic variation of neuraminidase stalk deletions in avian influenza viruses. PLoS One 6:e14722.

Li Y, Yamakita Y, Krug RM. 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 95:4864-4869.

Lu L, Lycett SJ, Leigh Brown AJ. 2014. Reassortment patterns of avian influenza virus internal segments among different subtypes. BMC Evol Biol 14:16.

Lu. J, Raghwani J, Wu J, Pryce R, Bowden T, Julien Thézé J, Huang S, Song Y, Zou L, Liang L, et al. 2018. Molecular evolution, diversity and adaptation of H7N9 influenza A viruses in China. Emerging Infectious Diseases 24.

Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley JL, Tong L. 2006. Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. Nature 444:953-956.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26:2462-2463.

Melen K, Kinnunen L, Fagerlund R, Ikonen N, Twu KY, Krug RM, Julkunen I. 2007. Nuclear and nucleolar targeting of influenza A virus NS1 protein: striking differences between different virus subtypes. J Virol 81:5995-6006.

Mok BW, Song W, Wang P, Tai H, Chen Y, Zheng M, Wen X, Lau SY, Wu WL, Matsumoto K, et al. 2012. The NS1 protein of influenza A virus interacts with cellular processing bodies and stress granules through RNA-associated protein 55 (RAP55) during virus infection. J Virol 86:12695-12707.

Monne I, Fusaro A, Nelson MI, Bonfanti L, Mulatti P, Hughes J, Murcia PR, Schivo A, Valastro V, Moreno A, et al. 2014. Emergence of a highly pathogenic avian influenza virus from a low-pathogenic progenitor. J Virol 88:4375-4388.

Munier S, Larcher T, Cormier-Aline F, Soubieux D, Su B, Guigand L, Labrosse B, Cherel Y, Quere P, Marc D, et al. 2010. A genetically engineered waterfowl influenza

virus with a deletion in the stalk of the neuraminidase has increased virulence for chickens. J Virol 84:940-952.

Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, et al. 2015. Gene-wide identification of episodic selection. Mol Biol Evol 32:1365-1371.
Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8:e1002764.

Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676-679.

Pond SL, Murrell B, Poon AF. 2012. Evolution of viral genomes: interplay between selection, recombination, and other forces. Methods Mol Biol 856:239-272.

Qi L, Pujanauski LM, Davis AS, Schwartzman LM, Chertow DS, Baxter D, Scherler K, Hartshorn KL, Slemons RD, Walters KA, et al. 2014. Contemporary avian influenza A virus subtype H1, H6, H7, H10, and H15 hemagglutinin genes encode a mammalian virulence factor similar to the 1918 pandemic virus H1 hemagglutinin. MBio 5:e02116.

Saenz RA, Essen SC, Brookes SM, Iqbal M, Wood JL, Grenfell BT, McCauley JW, Brown IH, Gog JR. 2012. Quantifying transmission of highly pathogenic and low pathogenicity H7N1 avian influenza in turkeys. PLoS One 7:e45059.

Schmolke M, Manicassamy B, Pena L, Sutton T, Hai R, Varga ZT, Hale BG, Steel J, Perez DR, Garcia-Sastre A. 2011. Differential contribution of PB1-F2 to the virulence of highly pathogenic H5N1 influenza A virus in mammalian and avian species. PLoS Pathog 7:e1002186.

She YM, Farnsworth A, Li X, Cyr TD. 2017. Topological N-glycosylation and site-specific N-glycan sulfation of influenza proteins in the highly expressed H1N1 candidate vaccines. Sci Rep 7:10232.

Short KR, Richard M, Verhagen JH, van Riel D, Schrauwen EJ, van den Brand JM, Manz B, Bodewes R, Herfst S. 2015. One health, multiple challenges: The inter-species transmission of influenza A virus. One Health 1:1-13.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol 32:1342-1353.

Sorrell EM, Song H, Pena L, Perez DR. 2010. A 27-amino-acid deletion in the neuraminidase stalk supports replication of an avian H2N2 influenza A virus in the respiratory tract of chickens. J Virol 84:11831-11840.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.

Stern DL. 2013. The genetic causes of convergent evolution. Nat Rev Genet 14:751-764.

Sun Y, Zhang Y, Hamilton K, Manley JL, Shi Y, Walz T, Tong L. 2018. Molecular basis for the recognition of the human AAUAAA polyadenylation signal. Proc Natl Acad Sci U S A 115:E1419-E1428.

Suzuki Y. 2006. Natural selection on the influenza virus genome. Mol Biol Evol 23:1902-1911.

Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. 1992. Evolution and ecology of influenza A viruses. Microbiol Rev 56:152-179.

Webster RG, Peiris M, Chen H, Guan Y. 2006. H5N1 outbreaks and enzootic influenza. Emerg Infect Dis 12:3-8.

Wikramaratna PS, Pybus OG, Gupta S. 2014. Contact between bird species of different lifespans can promote the emergence of highly pathogenic avian influenza strains. Proc Natl Acad Sci U S A 111:10767-10772.

Worobey M, Han GZ, Rambaut A. 2014. A synchronized global sweep of the internal genes of modern avian influenza virus. Nature 508:254-257.

Xu Y, Bailey E, Spackman E, Li T, Wang H, Long LP, Baroch JA, Cunningham FL, Lin X, Jarman RG, et al. 2016. Limited Antigenic Diversity in Contemporary H7 Avian-Origin Influenza A Viruses from North America. Sci Rep 6:20688.

Yang H, Chen LM, Carney PJ, Donis RO, Stevens J. 2010. Structures of receptor complexes of a North American H7N2 influenza hemagglutinin with a loop deletion in the receptor binding site. PLoS Pathog 6:e1001081.

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. 2015. The I-TASSER Suite: protein structure and function prediction. Nat Methods 12:7-8.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.