Parallel Evolution in the Emergence of Highly Pathogenic Avian Influenza A Viruses

Marina Escalera-Zamudio[1†*], Michael Golden[1†], Bernardo Gutiérrez[1], Julien Thézé[1], Jeremy Russell Keown[2], Loic Carrique[2], Thomas A. Bowden[2] and Oliver G. Pybus[1,3*]

1. Department of Zoology, Oxford University, Parks Rd Oxford, OX1 3PS, UK

2. Division of Structural Biology, Wellcome Centre for Human Genetics, Oxford, OX3 7BN, UK

3. Royal Veterinary College, London, UK.

[†]MEZ and MG contributed equally to this paper.
[*]Corresponding authors:
marina.escalerazamudio@zoo.ox.ac.uk; oliver.pybus@zoo.ox.ac.uk

other email addresses:
michael.golden@zoo.ox.ac.uk
bernardo.gutierrez@zoo.ox.ac.uk
julien.theze@zoo.ox.ac.uk
jeremy@strubi.ox.ac.uk
loic@strubi.ox.ac.uk
thomas.bowden@strubi.ox.ac.uk

ABSTRACT (145 words)

Parallel molecular evolution and adaptation are important phenomena commonly observed in viruses. Here we exploit parallel molecular evolution to understand virulence evolution in avian influenza viruses (AIV). Highly-pathogenic AIVs independently evolve from low-pathogenic ancestors via acquisition of a polybasic cleavage sites (pCS). Why some AIV lineages but not others evolve in this way is unknown. We hypothesise that the parallel emergence of highly-pathogenic AIV may be facilitated by permissive or compensatory mutations occurring across the AIV genome. We combined phylogenetic, statistical and structural approaches to discover parallel mutations in AIV genomes associated with the highly-pathogenic phenotype. Parallel mutations were screened using a new statistical test of mutation-phenotype association and further evaluated in the contexts of positive selection and protein function. The mutational panel we present reveals new links between virulence evolution and other viral traits and raises the possibility of predicting aspects of AIV evolution.

INTRODUCTION (5066 words)

Avian Influenza A viruses (AIVs) infect bird populations worldwide [1]. While most AIVs cause only a mild-to-asymptomatic infection, highly-pathogenic (HP) avian influenza, restricted to the H5 and H7 subtypes of AIV, can cause extremely high mortality (up to 100%) in domestic bird populations [1]. Both low-pathogenic (LP) and HP viruses can cause severe disease in humans following avian-to-human transmission, and there are potential instances of human-to-human transmission [2,3]. Recently, the emergence of H7N9 AIV in China caused >1500 infections and 615 deaths in people between 2013 and early 2018 [4]. Low virulence AIVs have evolved into HP strains on multiple occasions, resulting in geographically and temporally distinct HP outbreaks that generate sizeable economic losses to the poultry industry, and pose an increased risk for avian-to-human spread [5 6]. Therefore, it is important to understand the mechanisms driving the evolution of virulence in AIVs.

The virulence of an infectious disease is, in general, a complex trait [7 8 9] . However, the AIV HP phenotype is one of the simplest and best characterized molecular markers of virulence, and is therefore an attractive system in which to study virulence evolution. The HP phenotype is conferred by a polybasic cleavage site (pCS) within the virus haemagglutinin (HA) protein. A pCS is the outcome of one or more insertion mutations resulting in multiple basic amino acids, which renders the HA protein a target for broadly-expressed host proteases, facilitating systemic virus spread within avian hosts [10 11]. The pCS insertion has occurred several times during the evolutionary history of AIVs. Over ten different pCS sequences have been described for HP H5NX viruses since 1961, and over fifteen pCS have been described for HP H7NX viruses since 1963 [12]. The evolutionary pathways that lead to the independent, parallel evolution of pCSs and the emergence of the HP phenotype in different AIV lineages are not well understood.

Parallel molecular evolution describes repeated evolutionary changes leading to the same phenotype or genotype in independent populations, and can result from adaptation by natural selection to similar selective pressures or ecological niches [13,14]. RNA viruses exhibit

high rates of mutation and replication, large population sizes, consistently varying environments, and strong selective pressures. These factors, combined with small genome sizes, which may restrict the number of mutational pathways leading to a given genotype, may result in parallel molecular evolution being common in RNA viruses [15].

We hypothesize that the parallel evolution of the HP genotype from LP ancestors in AIV subtypes H5 and H7 may have a genetic basis, and may be therefore associated with permissive or compensatory secondary mutations occurring elsewhere in the virus genome. Such mutations could arise prior to or immediately after the emergence of a pCS, perhaps to compensate for possible detrimental effects on the stability of the HA protein [16,17]. Little is known about the genetic predisposition of AIV to acquire a pCS (except for RNA structural studies of the cleavage site region itself [18]), nor about the role of mutations elsewhere in the virus genome favouring the evolution to a HP phenotype.

Here we use a combination of analytical methods from evolutionary and structural biology to detect parallel mutations across the AIV genome that are associated with the evolution of the HP phenotype (hereafter termed *HP-cluster associated parallel mutations, or HAPMs*). We developed, validated and implemented a new phylogenetic model of binary trait evolution that tests whether genotype-to-phenotype associations observed across multiple lineages are statistically significant. This approach explicitly incorporates the phylogenetic correlation among observations and thus is less susceptible to error than standard statistical tests [19]. We further investigated if the HAPMs detected had evolved under positive selection, and assessed their possible effects on protein function and/or stability. We identify numerous HAPMs among HP AIV lineages of same viral subtype, many of which are also associated with other viral processes, including antigenic escape, host-response modulation, and replication efficiency. We discuss how the mutational panel presented here can provide a starting point for future molecular studies on the evolution of the HP phenotype.

METHODS (1524)

We developed a new analysis strategy to detect parallel genetic changes associated with the LP-to-HP evolution of AIV (summarized in Figure 1). For convenience, we refer to parallel genetic changes as HAPMs (*HP-cluster associated parallel mutations*). The list of HAPMs identified for each virus subtype was further analysed to identify those most likely to be functionally and selectively relevant using the following criteria: *i)* they exhibit a statistically-supported association with the HP phenotype, *ii*) they show evidence of being evolving under positive selection, *iii)* they are predicted to be stabilizing to protein structure, and *iv)* they have known or predicted functional relevance, derived from structural analysis.

1. Large-scale phylogenetic analysis

All available coding sequences for all genome segments of H7NX and H5NX viruses, from all hosts and geographical regions, were downloaded from the Influenza Virus Database. Sequences were excluded if (i) they were >200nt shorter than the full coding sequence length, (ii) they represented identical sequences, or (iii) they contained >10% ambiguities. Sequences were aligned using MAFFT [20] and preliminary large-scale phylogenies for each genome segment were estimated using RAxML under a general time reversible nucleotide substitution model with gamma distributed among-site rate variation (GTR+G). Branch support was assessed using a bootstrap approach with 100 replicates [21]. HP strains were identified by (i) detecting >1 basic residue insertions within the consensus cleavage site in the HA protein alignments [12] and (ii) by confirming HP status with case reports in the literature [5]. Phylogenetic clusters of HP sequences were identified as (i) monophyletic and well supported (bootstrap value >80) clusters that, (ii) contain >2 HP sequences, (iii) descend from LP ancestors in all genome segments, and (iii) have emerged independently from other HP clusters, and are thus distinct in time and space.

2. Phylogenetically-informed subsampling

To render further analyses computationally feasible, we sub-sampled the abovementioned datasets in a phylogenetically-informed manner. Subsampling was undertaken so as to retain the mutational pathways along internal tree branches that lead from LP ancestors to HP clusters, whilst reducing overall alignment sizes. For each genome segment, we retained (i) selected HP clusters, (ii) their immediate LP ancestors in the tree, (iii) up to 250 randomly-selected LP reference sequences, in order to maintain overall tree topology, and (iv) a group of ancestral sequences for rooting purposes. Sub-sampled datasets were re-aligned and screened for recombination using the RDP, GENECOV and BOOTSCAN components of the RDP3 package [22]. No evidence for recombination was detected by three or more methods under a Bonferroni corrected *p*-value<0.001. Prior to estimation of final phylogenies, HA alignments were edited to remove the cleavage site and NA alignments were edited to remove hypervariable regions with poor alignment confidence. Trees were estimated from the sub-sampled alignments using RAxML under the GTR+G model and bootstrapped with 100 replicates [21]. Accession numbers and alignment sizes are provided in Supplementary File 1. The varying sizes and/or missing HP clusters in some alignments are due to the lack of complete genomes for all strains in the public database.

3. Detecting HAPMs occurring across HP clusters

To screen for HAPMs within AIV genomes, all mutations (deletions, insertions and non-synonymous variable sites) present in ≥1% of the sequences in each sub-sampled alignment were coded as a discrete taxon trait and subjected to ancestral state reconstruction, using the maximum likelihood framework implemented in the "*ape*" R package [23]. The resulting phylogenetic reconstructions were inspected visually to identify HAPMs shared by multiple HP clusters. HAPMs were identified if: (i) they occurred within 2 or more HP clusters, (ii) they were present in >60% of the sequences within the HP clusters, and (iii) they were not observed on internal branches anywhere else in the phylogeny (Supplementary File 2).

127 To illustrate in detail the evolutionary association between HAPMs and AIV HP
128 clusters, we selected three HAPMS for each virus subtype for ancestral state reconstruction
129 using Bayesian molecular clock phylogenetic analysis. These six examples were chosen to
130 illustrate the range of evolutionary dynamics exhibited by HAPMs, and because they are all
131 significantly associated with the HP phenotype (see next section). For each site, we
132 estimated the posterior probability of amino acid states at tree nodes ancestral to HP clusters
133 in the subsampled alignments, using the discrete trait evolution model in BEAST v1.8.4 [24].
134 Time-scaled phylogenies were estimated under a SRD06 nucleotide substitution model, a
135 strict molecular clock model and a Bayesian Skyline coalescent tree prior. Partitions were
136 generated for each site, and ancestral states/transition rates were estimated for each
137 mutation using an asymmetric continuous time Markov chain model [25,26]. Two MCMC runs
138 were computed for 100×10^6 states or until convergence was reached. Maximum clade
139 credibility (MCC) trees were summarized using TreeAnnotator, after removing 10% of the
140 runs as burn-in [24]. Additionally, we re-estimated the full evolutionary history of these HAPMs
141 by mapping the amino acid changes onto the large-scale ML phylogenies generated in step 1
142 (Supplementary File 3).
143
144 **4. Testing for genotype-to-phenotype associations**
145 We developed a new evolutionary model to statistically test for associations between
146 binary phenotypic traits and amino acid states in a phylogenetically-informed manner. We
147 used this model to determine which of the HAPMs identified above were significantly
148 associated with the HP phenotype. Our method serves the same purpose as a χ^2 test, but
149 correctly incorporates the phylogenetic history of the trait and phenotype of interest, as it is
150 known that failing to account for these phylogenetic correlations can lead to erroneous false
151 positives [19,27]. The model assumes that phenotypes and amino acid states are discrete traits
152 that change along phylogeny branches. Given an alignment, its phylogeny, and a set of
153 binary values for each sequence, the method proceeds to test if a specific amino acid state is
154 associated with the given phenotype. This is repeated for all amino acid states and all

155 codons. A coupling parameter λ is used to measure the strength of association between an
156 amino acid state and the phenotype state. The null hypothesis of zero association is
157 represented by $\lambda = 1$, and can be rejected in favour of a positive association (λ >1) using a
158 likelihood ratio test (LRT). The type I and II error rates of this test were evaluated using
159 simulations. A full mathematical description of the model and simulation details are provided
160 in Supplementary Text 5. We applied this model to our subsampled H5NX and H7NX
161 alignments. The H5NX alignments were further reduced by collapsing clade C8 to only its
162 most basal sequences, in order to avoid the overrepresentation of HP sequences.
163
164 **5. Testing for diversifying positive selection**
165 To identify positively-selected HAPMs, we used various *dN/dS* (ω) ratio estimation
166 methods. The branch-site model A (BSA) in CODEML [28] was used to identify positively-
167 selected sites (PSS) associated with specific phylogeny branches, namely the HP clusters.
168 The statistical significance of the BSA model was compared to a null hypothesis with an ω=1
169 and evaluated using a LRT. The datamonkey server [29] was also used to test for sites
170 evolving under pervasive and episodic positive selection under both site- (MEME: Mixed
171 Effects Model of Evolution; SLAC: Single-Likelihood Ancestor Counting and FEL: Fixed
172 Effects Likelihood) and branch-site models (BSREL: branch-site random effects likelihood)
173 [30,31,32,33]. Test significances were assessed using a LRT and subjected to Holm-Bonferroni
174 multiple test correction. Finally, PSS were compared to the HAPMs detected and to the
175 DEPS algorithm results (Directional Evolution of Protein Sequences using Maximum
176 Likelihood) [34] (Supplementary Table 1). The LRT results and lists of the PSSs identified are
177 provided in the Supplementary File 4.
178
179 **6. Inferring mutation effects on protein stability and function**
180 We undertook structural analyses to assess if the identified HAPMs might be
181 functionally relevant. To determine the location and potential functional implications of the

182 different HAPMs, we modelled protein 3D structures from existing structural data in the PDB
183 database (https://pdb101.rcsb.org/), using representative HP sequences from our datasets
184 as queries. For H7, we used the crystal structure of A/H7N2/New York/107/2003 AIV (PDB
185 ID:3M5G), whilst for H5 we used the crystal structure of A/Vietnam/30262III/04 AIV (PDB ID:
186 2IBX). The high-resolution structure PDB 2Q06 was used to model NS1, and PDB 2Z16 was
187 used to model M1. For the PB2/PB1/PA complex, the high-resolution structure from the bat
188 Influenza A/H17N10 virus (PDB 4WSB) was used as a template, as this is the only structural
189 data currently available for the complete influenza A virus polymerase complex [35]. For the
190 HA, M1, and NS1 proteins, we used the I-TASSER server [36] to generate models with the
191 mutations of interest, that were compared manually to the PDB files for validation. For the
192 polymerase complex, mutations were introduced to the superimposed 3D protein structures
193 using PyMOL (Schrödinger, LLC). Finally, to predict which HAPMs had a possible effect on
194 protein stability, we used the PIPS (Phylogenetic Inference of Protein Stability) program [37].
195 This program implements an informed Bayesian approach and estimates, for each amino
196 acid state, a value of free energy change (ddG) in terms of the thermodynamic stability of an
197 unfolding/refolding protein, weighted under a phylogenetic background. Negative ddG values
198 are predicted to be stabilizing for protein structure. Values close to zero are predicted to have
199 little effect or to be only weakly stabilizing, whilst positive values are predicted to destabilize
200 protein structure [37].

201 **RESULTS**

202

203 **Large-scale phylogenetic analysis of independent HP clusters**

204 Large-scale phylogenetic analysis confirmed that H7NX viruses are split into two
205 major clades, the Eurasian and North American lineages (Supplementary Figure 1) [38]. H7 HP
206 outbreaks occurred independently through time and space and correspond to distinct
207 lineages that, in most cases, display a unique pCS motif [5,6] (Figure 2). LP H7NX viruses
208 carry a consensus amino acid sequence -PEXPKGRGLF- at the HA cleavage site, whereas
209 a minimal consensus of -PEXPKXnK/RnRGLF- is identified for HP viruses (Figure 2a) [39] [11].
210 Nine H7NX well-supported clusters and spatio-temporally distinct HP lineages were identified
211 in the HA phylogeny (Figure 2b) and in trees of other genome segments (Supplementary
212 Figs. 2,3). Minor inconsistencies were observed for cluster C1, which segregated into C1.1
213 and C1.2 in the PB2, PB1, PA, M, and NS1 trees (Supplementary Figure 2).

214 The large-scale phylogeny of H5NX viruses shows a basal clade of HP and LP H5NX
215 viruses circulating in wild birds worldwide from the late 1960s to the early 2000s (Figure 2d,
216 Supplementary Figure 1). This basal clade is ancestral to the large and long-lived HP H5N1
217 lineage (named C8 here). This clade emerged in 2003 in China and subsequently spread to
218 Asia, Africa, Europe and America [1] and persists until today [5]. Within C8, there is a distinct
219 group of HP H5NX viruses (H5N2, H5N6 and H5N8 subtypes) that circulates in both wild and
220 domestic birds (Figure 2d). Eight H5NX HP clusters were identified as being spatio-
221 temporally distinct and well-supported (Figure 2c). Most HP clusters displayed a unique pCS
222 sequence, except for C8, reflecting the ongoing evolution of this long-lived lineage. As with
223 the H7NX viruses, phylogenetic clustering of HP H5NX viruses was conserved within other
224 genome segment phylogenies (Supplementary Figures 3,4).

225

226 **Identifying HAPMs significantly associated with the HP phenotype**

227 The number of amino acid sites that were polymorphic (i.e. variants present at >1%
228 frequency) varied among virus subtypes and genome segments. For H7NX viruses, the

229 number of polymorphic sites per segment ranged from 21 (for M1) to 125 (for HA). Maximum
230 likelihood ancestral state reconstructions revealed 16 parallel genetic changes (HAPMs)
231 present in two or more H7NX HP lineages (Supplementary File 2). For H5NX viruses,
232 between 12 (for M1) and 141 (for HA) polymorphic sites were detected for each segment.
233 We identified 39 HAPMs present in two or more H5NX HP lineages (Supplementary File 2).
234 Of these, 17 occurred within sub-lineages of cluster C8 and thus were not considered further.
235 To focus attention on those HAPMs significantly associated with the HP phenotype,
236 we developed a new phylogenetic model to test the null hypothesis of no association
237 between mutations at each position and the HP phenotype (Supplementary Text 5). For
238 H7NX viruses, we found a strong association with the HP phenotype for 9 out of 16 HAPMs
239 (Table 1). For H5NX viruses, a strong association with the HP phenotype was supported for
240 12 out of 22 HAPMs (Table 1).
241 To illustrate the evolutionary dynamics of HAPMs, we chose a subset of six HAPMs
242 that were significantly associated with the HP phenotype and subjected them to molecular
243 clock phylogenetic analysis (Figure 3). We observed a variety of evolutionary patterns and in
244 general obtained high posterior probability values for ancestral amino acids at nodes
245 preceding the emergence of HP clusters. By way of illustration, consider site 355 of H7NX
246 PB2. At this site K is present in HP clusters C5 and C6, but not in lineages representing the
247 LP background (Figure 3a). A similar pattern is seen for mutation 627K in H5NX viruses
248 (Figure 3f). More extensive parallel evolution is exhibited by mutations 143T in H7 (Figure
249 3b) and 127L in H5 (Figure 3d), which are shared by multiple HP clusters, but again not
250 observed within LP lineages. Mutation 209N in H7NX NS1 is found within several HP
251 clusters, but is also common in LP sequences of the American lineage (Figure 3c). Lastly,
252 site 154 in H5 exhibits a more complex, step-wise evolutionary process, from N to I/L to Q,
253 reflecting the transition from LP to HP (Figure 3e). N is present only in LP viruses, I/L occurs
254 in HP viruses circulating prior to and during the early expansion of C8, and Q is only present
255 in more recent HP viruses from C8 (Figure 3e).

256 To facilitate computational analyses, we sub-sampled AIV alignments before
257 detecting HAPMs (see Method step 2). Thus, the original alignments before subsampling
258 may contain HP lineages that were not present after sub-sampling (Supplementary Figure 1).
259 Thus, we investigated whether if these lineages also contained any HAPMs significantly
260 associated with the HP phenotype, by mapping HAPM amino acid changes onto the full
261 phylogenetic trees (Supplementary File 3). Since these HP lineages were not used to define
262 HAPMs, mutations present in them constitute an independent test of our statistical approach.
263 We found that, in addition to the HP clusters used here, mutation 154L in H5 was fixed in two
264 independent HP H5N1 outbreaks in Vietnam from 2008-2013 [40], and in HP H5N2 viruses in
265 China circulating from 2009-2015 [41] (Supplementary File 3). We also observed that mutation
266 143T is present in H7N6 viruses isolated from an outbreak in Aichi prefecture, Japan, in 2009
267 [42] (Figure 3b; Supplementary Files 1,3). Although this was not a HP outbreak, these viruses
268 contained an additional arginine residue within the pCS (-PKGR to -PKRR), which triggered
269 concerns about whether this lineage might evolve the HP phenotype. Fortunately, this
270 lineage was successfully eradicated before any HP phenotype strains were observed [42].
271 Most HAPMs were not observed on phylogeny branches immediately ancestral to the
272 HP clusters, suggesting that these appear shortly before, during, or after the emergence of
273 the HP phenotype and making it difficult to discern if they are permissive or compensatory
274 change. However mutation A143T in H7 is present in LP viral sequences that are basal and
275 immediately ancestral to the emergence of the HP H7N1 outbreak in Italy 2000 [43]. This
276 observation supports the notion that at least some of the HAPMs identified here could be
277 permissive.
278
279 **HAPMs associated with diversifying positive selection**
280 Although parallel evolution is often caused by positive selection, parallel genetic
281 changes can, in theory, also arise by chance [14]. Consequently, we investigated whether if
282 some HAPMs were inferred as being positively selected sites (PSS; see Methods section 5).
283 Analyses under site- and branch-site models showed evidence for diversifying positive

284 selection at some HAPMs, and for some HP clusters (H7NX clusters C4 and C7, and H5NX
285 clusters C1, C2 and C4; Supplementary Table 1). As expected [15], *dN/dS* (ω) values for whole
286 genome segments were <1 in all cases (data not shown), and the overall proportion of PSS
287 detected was low (Supplementary Table 1, Supplementary File 4).

288 For the H7NX viruses, 4/16 HAMPs (corresponding to 4/9 of the HAPMs significantly
289 associated with the HP phenotype) were identified as PSS. For the H5NX viruses, only 2/22
290 HAPMs (corresponding to 2/12 of the HAPMs significantly associated with the HP
291 phenotype) were identified as PSS (see Supplementary Table 1 for details). At first
292 impression, this suggests that most HAPMs significantly associated with the HP phenotype
293 are not a direct consequence of positive selection, and may instead represent neutral
294 variation that has arisen through genetic hitchhiking [14,16]. However, if some HAPMs are
295 selectively neutral in the absence of a pCS (LP phenotype), but offer a positive effect when a
296 pCS is present (HP phenotype), then they may not be inferred as PSS using *dN/dS* methods
297 yet still represent a permissive/compensatory change. This is a consequence of *dN/dS*
298 estimation methods being conservative in detecting sites under positive selection in some
299 circumstances [44]. For example, mutation E627K in PB2, which was not inferred to be a PSS,
300 is known to have a positive effect on the adaptation of AIV to mammalian hosts and
301 increases replication efficacy [45]. The role of selection on whole HP lineages is less clear: for
302 example, experiments in turkey populations suggest that the HP and LP phenotypes of H7N1
303 viruses have indistinguishable transmission rates and basic reproductive numbers (R_0) [46].
304

305 **Structural and protein stability analyses**

306 Most of the HAPMs significantly associated with the HP phenotype had negative ddG
307 values, as assessed under a Bayesian approach for inferring the stability effects of individual
308 mutations from homologous protein sequences with a known phylogeny [37] (Figures 4 and 5).
309 Thus, they are predicted to be fairly stabilizing/neutral changes to protein structure. One
310 exception is E627K in H5NX PB2, which is predicted to be destabilizing. However, it has

311 been shown that other compensatory mutations in PB2 appear to improve the genetic
312 stability of E627K [47].

313 Structural analysis suggests that several H7NX HAPMs significantly associated with
314 the HP are functionally relevant (Figure 4). A143T in HA is a non-conservative amino acid
315 change located in a solvent-accessible region within the 130-loop proximal to the receptor
316 surface. This is within antigenic pocket A, contributing to the electrostatic environment
317 needed for ligand binding. Although it has no direct interaction with glycans [48], mutation
318 A143T induces glycosylation of residue 141 by adding a N-linked glycosylation motif. This N-
319 linked residue produces steric masking of the antigenic site and may affect HA folding and
320 stability [75]**.** Similarly, M/L271I in HA is located within the receptor binding subdomain [49,50].
321 R355K in PB2 is a conservative amino acid change within a solvent-exposed loop of the
322 electrostatic surface of the cap binding domain. This site is in direct contact with first
323 phosphate after the cap [51] and correlates with high pathogenicity of AIVs in mouse models [52].
324 D209N in NS1 is a conservative amino acid change located within a disordered tail
325 interacting with importin alpha, a cellular host factor that promotes the entry of proteins to the
326 cell nucleus (Figure 4), and may thus interfere with cell cycle regulation pathways [53].

327 HAMPs significantly associated with the HP phenotype in H5NX viruses were also
328 found to be structurally relevant (Figure 5). Mutation F127L is a conservative amino acid
329 change within the globular head of HA1, close to antigenic site 3 [54]. N154I/L is a non-
330 conservative amino acid change within antigenic site 2, and involves the loss of the
331 asparagine-linked glycosylation site [55]. Both these HAPMs correspond to variable sites
332 related to escape mutants [54]. However, they are not crucial for the protective efficacy of H5
333 vaccines, suggesting that the loss of the saccharide linkage may have little impact on
334 antigenic variation [50,55]. I167T is located in the surface-exposed region of the HA protein, and
335 has been related to increased infectivity in the human lower respiratory tract in HP H5N1
336 viruses from Egypt [56]. S145L in HA is also located in the surface exposed region and has
337 been linked to increased virulence in poultry [57]. Position 145 also undergoes occasional
338 deletion (as observed in approximately 10% of sequences) and it has been argued that this

deletion is critical for the antigenic evolution of long-lived HPAIV clusters [75]. Mutation E627K in PB2 is a well-characterised adaptation marker associated with enhanced transmission in mammalian hosts [58], found frequently among H5NX viruses causing human infections [7]. Site 627 is surface exposed and located on the “627 domain” and is essential for viral RNA replication and transcription in a cellular context [59] (Figure 5). Mutation A674T in PB2 is located at the interface of the polymerase complex trimer and is predicted to increase strength of the PB2 protein structure. In *in vivo* mouse models this mutation is a prerequisite for the acquisition of E627K, as it stabilizes polymerase activity when co-occurring with 627K [60], supporting the notion that this HAPM acts as a permissive mutation.

348 **DISCUSSION**

349 The occurrence of parallel molecular evolution in AIV genomes can be used as a
350 starting point for the identification of mutations associated with the evolution of the HP
351 phenotype. We report combined evolutionary and structural evidence to support the
352 hypothesis that some AIV parallel mutations are significantly associated with the HP
353 phenotype and may potentially act as permissive/compensatory changes to facilitate the
354 evolution of HP viruses. Previous studies have noted parallel genetic changes in
355 independent AIV lineages [61], but these were not linked to the emergence of the HP
356 phenotype. Similarly, specific instances of parallel genetic changes associated with the
357 transition from a LP to a HP phenotype giving rise to single H7N1 outbreaks in Italy have
358 also been described [43,62,63]. We provide the first systematic and statistically-rigorous
359 screening of AIV mutations associated with HP outbreaks. We demonstrate the effectiveness
360 of phylogenetically-informed hypothesis testing coupled with comparative evolutionary and
361 structural approaches to identify potentially important mutations exhibiting parallel evolution
362 in H5 and H7 AIV genomes.

363 Permissive/compensatory mutations associated with the HP phenotype are expected
364 to be favoured under a HP genetic background, and therefore occur repeatedly in HP
365 lineages. Mutations that occur at nodes preceding and coinciding with the emergence of
366 pCS-defined HP clusters (*e.g.* those that are present in the immediate LP ancestors and in
367 the HP cluster), may represent permissive changes that facilitate the subsequent evolution of
368 the HP phenotype. Conversely, mutations that occur only within HP clusters (after a pCS has
369 arisen), may be more likely to be compensatory changes.

370 However, sampling of AIV genomes is likely biased towards HP outbreaks, resulting
371 in a phylogenetic under-representation of the immediate LP ancestors of any given HP
372 lineage. This under-sampling of local LP strains makes it difficult to discern whether the
373 HAPMs correspond to permissive versus compensatory changes. Further molecular
374 experiments are needed to understand how the parallel mutations detected here actually
375 contribute to the evolution of the HP phenotype by, for example, stabilizing protein structure,

altering host-virus or virus-virus inter-molecular interactions, or by combining additively to increase virulence. However, such experimental validation may take several years to complete especially if HP AIV gain-of-function experiments will be required [64]. In the meantime, we should continue to exploit fully the information revealed by parallel molecular evolution, which represents a "natural" replicated experiment. This situation further highlights the importance of systematic sampling and whole genome sequencing of both HP viruses *and* their immediate LP ancestors, prior and during AIV outbreaks.

Some of the HAPMs identified here were previously identified either *in vitro* or *in silico* as being associated with HP AIV (Table 1) [47,52,56-58,60] . Although this supports our results, it is possible that the number, biological roles, and genomic distribution of mutations involved in HP adaptation is currently underestimated. Notably, the HAPMs described here are not shared among the H7NX and H5NX virus subtypes, yet they commonly occur within the same genome segments (Table 1), suggesting that the evolution of the HP phenotype may interact with other well-studied biological processes, such as antigenic escape (determined by the HA protein), host-response modulation (regulated by PB2 and NS1) and replication efficiency (PB2). Our findings are in agreement with previous studies showing that the acquisition of a pCS may require coadaptation of the HA and other viral proteins [65].

Some of the mutations occurring close to antigenic sites may be indirectly associated with differential patterns of immunity in varied host populations, defined by complex factors such as animal age and sex distributions, previous exposure to infections/vaccines, or differences in host genetic diversity [66]. Interestingly, mathematical modelling has suggested that the sporadic population dynamics of HP strains may be determined, at least in part, by their linkage to one or more specific antigenic variants, whose frequencies are driven by the dynamics of host immunity, rather than by mutability [67,68]. Furthermore, the emergence of the HP phenotype has been often associated with virus adaptation from wild to domestic bird host populations [61]. The strength of this association is difficult to assess in nature given the substantial bias (in both the literature and sequence databases) towards AIV infections and HP outbreaks in poultry versus those in wild birds. However, using our new model of

404 mutation-trait association (Methods section 4; Supplementary Text 5), we found no
405 association between any of the detected HAMPs (or any other mutation) and the host
406 species (domestic versus wild bird; data not shown). This suggests that the parallel
407 mutations identified here are not directly linked to transmission among domestic bird
408 populations, although high population densities together with a limited genetic variability
409 within the host (as occurs in intense farming facilities), may facilitate virus transmission. If
410 acquiring a pCS does indeed reduce AIV transmission fitness in natural systems [46], then it is
411 interesting to speculate that both intensive farming and compensatory/permissive mutations
412 may act to facilitate the spread of a HP strain; the latter by increasing R_0 and the former by
413 lowering the threshold R_0 required to sustain transmission in a population.

414 Lastly, we propose that the set of HAPMs associated with HPAIV lineages identified
415 here (or any improvement to this set obtained in the future), may, with sufficient validation
416 and testing, function as an early detection system capable of highlighting lineages that pose
417 a greater risk of evolving the HP phenotype in the future. Importantly, the mutational panel
418 identified here could potentially benefit to AIV control and bird health even if the mechanistic
419 and functional actions of the mutations described are unknown. We suggest that circulating
420 AIV strains in domestic and wild bird populations that do not contain a pCS site, but which
421 exhibit a combination of the HAPMs identified here, are potential targets for more focussed
422 and targeted genomic and epidemiological surveillance.

423

424 **AUTHOR CONTRIBUTIONS**

425 MEZ and OGP designed research. MEZ performed research. MEZ and OGP analysed data.
426 MG developed the statistical model and analysed data. BG and JT analysed genomic data.
427 JRK and LC analysed structural data. TAB supervised structural data analysis. MEZ wrote
428 the manuscript, with comments from all authors.

429

430 **ACKNOWLEDGEMENTS**

MEZ is supported by an EMBO Long Term Fellowship (ALTF376-2017). MG, JT and OGP are supported by the EU's Seventh Framework Programme (FP7/2007-2013)/European Research Council (614725-PATHPHYLODYN). BG is supported by the 2017 Universities of Academic Excellence Scholarship Program of the Secretariat for Higher Education, Science, Technology and Innovation of the Republic of Ecuador (ARSEQ-BEC-003163-2017) and by Universidad San Francisco de Quito. JRK and LC are supported by the Wellcome Investigator award (200835/Z/16/Z). The Wellcome Centre for Human Genetics is supported by grant 203141/Z/16/Z. TAB is supported by the Medical Research Council (MR/L009528/1).

FIGURE LEGENDS

Figure 1. Summary of analytical approach

A summary of our approach to identifying HAPMs (HP-cluster associated parallel mutations) in AIV genomes. 1. Large-scale data set curation and phylogenetic analysis to detect genotypes and phenotypes of interest (HP lineages highlighted in red). 2. Phylogenetically-informed subsampling to reduce alignment size whilst retaining key mutational pathways along internal branches. 3. Detecting HAPMs across multiple HP clusters using ancestral state reconstructions (HAPMs shared among clusters represented by coloured dots) 4. Statistical testing for genotype-to-phenotype associations. 5. Identifying HAPMs evolving under positive selection using different *dN/dS* (ω) estimation methods. 6. Inferring the possible effects of HAPMs on protein stability and function by inferring ddG values, and by using existing structural data to infer interactions.

Figure 2. Independent spatio-temporal emergence of the HP phenotype

(a) Geographical occurrence of H7NX HP outbreaks. The year of circulation, virus subtype and consensus sequence for the polybasic cleavage site (pCS) are shown for the HP clusters analysed in this work. Countries of emergence are in red. (b) Phylogenetic distribution of HP outbreaks in the H7NX phylogeny. Distinct well-supported HP clusters within the HA tree are highlighted in red and named (C1-C9). Posterior probabilities and bootstrap support values are shown for nodes of interest. (c) Geographical occurrence of H5NX HP outbreaks, excluding C8 (clade 2.2) which is distributed worldwide. (d) Phylogenetic distribution of HP outbreaks in the H5NX phylogeny. Distinct well-supported HP clusters within the HA tree are highlighted in red and named (C1-C8). The H5NX HP viruses within cluster C8 (clade 2.2) are highlighted in blue.

Figure 3. Time-scaled reconstruction of amino acid evolution at selected HAPMs

Maximum clade credibility trees are shown for six illustrative HAPMs (a-f) that exhibit a significant associated with the HP phenotype. Amino acid changes at each of the six sites

469 were inferred using a Bayesian molecular clock phylogenetic approach. Branches are
470 coloured according to the inferred amino acid states at internal nodes (tip states not shown).
471 Ancestral nodes preceding the emergence of a given HAPM are highlighted by coloured
472 circles. The posterior probabilities for the amino acid state occurring at specified node are
473 shown. The HP clusters are highlighted by grey circles.
474
475 **Figure 4. Functional relevance of HAPMs of H7NX viruses**
476 Protein structure models for the HA (Hemagglutinin), PB2/PB1 (Polymerase complex) and
477 the NS1 (Non-structural protein) of the H7NX viruses. Functionally relevant HAPMs are
478 highlighted in red within the structural models; only those HAMPs that are significantly
479 associated with the HP phenotype are indicated. Estimated values for free energy change
480 (ddG) are indicated for these sites. Negative values or those close to zero can be interpreted
481 as fairly stabilizing/neutral mutations with respect to protein structure.
482
483 **Figure 5. Functional relevance of HAPMs of H5NX viruses**
484 Protein structure models for the HA (Hemagglutinin) and the PB2 polymerase subunit of the
485 H5NX viruses. See Figure 4 legend for further details.

486 **TABLE 1. HAPMs significantly associated with the HP phenotype of H5 and H7 virus subtypes**
487

H7NX							
HAPM	**Segment**	**HP clusters**	**Num. HP clusters**	**λ [95% CI]**	**p-value**	**H7 numbering**	**Reference***
A143T †	HA	C2, C5, C6, C7 and C8	5	7.1 [3.3, 13.2]	0.00	125	67,68,69,75
D209N †	NS1	C1.1, C6 and C8	3	2.4 [1.0, 4.3]	0.03		
R355K †	PB2	C5 and C6	2	6.0 [2.8, 11.8]	0.00		57
M/L274I	HA	C4 and C8	2	6.2 [3.1, 11.5]	0.00	256	
V480I	PB2	C5 and C6	2	9.2 [4.3, 18.9]	0.00		
V438I	HA	C3 and C8	2	4.6 [2.4, 8.3]	0.01	423	
D384N	HA	C6 and C8	2	5.1 [2.6, 9.5]	0.02	369	
G154D	PB1	C2 and C6	2	7.7 [4.3, 13.2]	0.03		67
S152L	PB1	C6 and C9	2	3.1 [1.7, 5.2]	0.03		
H5NX							
HAPM	**Segment**	**HP clusters**	**Num. HP clusters**	**λ [95% CI]**	**p-value**	**H5 numbering**	**Reference***
V113I	PB1	C2, C4 and C8	3	7.3 [3.7, 14.5]	0.01		
F127L†	HA	C3, C7 and C8	3	6.5 [3.3, 12]	0.03	111	59
N154I/L/Q† §	HA	C6, C7 and C8	3	See footnote	0.02	138	59
S145L/del‡	HA	C5 and C8	2	See footnote	0.01	129	62,75
A674T	PB2	C2 and C6	2	7.4 [3.9, 13.6]	0.00		65
M242I	HA	C1 and C8	2	7.4 [3.7, 13.7]	0.00	226	72
I167T	HA	C6 and C8	2	5.8 [3.0, 10.4]	0.03	151	61
E627K	PB2	C1 and C8	2	5.1 [2.7, 9.2]	0.04		
I451T	PB2	C2 and C8	2	3.8 [2.1, 6.8]	0.03		
R101K	M	C5 and C8	2	3.7 [1.7, 6.9]	0.02		
T204I	HA	C2 and C8	2	2.7 [1.4, 4.7]	0.03	188	
K379R	HA	C1 and C8	2	14.8 [7.4, 27.5]	0.00	363	

†Scored as positively selected site under different methods (see Supplementary Table 1 in the Supplementary Information).

§ Individual lambda values for each site are as follow: for 154I 1.7 [1.0, 2.7], for 154L 3.2 [1.9, 5.2] and for 154Q: 3.4 [2.0, 5.8], all with a p-value of 0.00.

‡ Individual lambda values for each site are as follow: for 145L 77.5 [69.5, 100.0] with a p-value of 0.00, and for 145del 7.0 [0.5, 14.7] with a p-value of 0.04.

* Where such HAPMs had been previously mentioned.

488

REFERENCES

1 Webster, R. G., Peiris, M., Chen, H. & Guan, Y. H5N1 outbreaks and enzootic influenza. *Emerg Infect Dis* **12**, 3-8, doi:10.3201/eid1201.051024 (2006).

2 Buxton Bridges, C. *et al.* 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, doi:10.1086/315213 (2000).

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

4 Lam, T. T. & Pybus, O. G. Genomic surveillance of avian-origin influenza A viruses causing human disease. *Genome Med* **10**, 50, doi:10.1186/s13073-018-0560-3 (2018).

5 Dhingra, M. S. *et al.* 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, doi:10.3389/fvets.2018.00084 (2018).

6 Short, K. R. *et al.* One health, multiple challenges: The inter-species transmission of influenza A virus. *One Health* **1**, 1-13, doi:10.1016/j.onehlt.2015.03.001 (2015).

7 Hatta, M., Gao, P., Halfmann, P. & Kawaoka, Y. Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science* **293**, 1840-1842, doi:10.1126/science.1062882 (2001).

8 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).

9 Schmolke, M. *et al.* Differential contribution of PB1-F2 to the virulence of highly pathogenic H5N1 influenza A virus in mammalian and avian species. *PLoS Pathog* **7**, e1002186, doi:10.1371/journal.ppat.1002186 (2011).

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

11 Abdelwhab, E. M. *et al.* Composition of the Hemagglutinin Polybasic Proteolytic Cleavage Motif Mediates Variable Virulence of H7N7 Avian Influenza Viruses. *Sci Rep* **6**, 39505, doi:10.1038/srep39505 (2016).

12 OFFLU. *OIE/FAO Network of expertise on animal influenza: Influenza A Cleavage Sites Report 2018*, <http://www.offlu.net/fileadmin/home/en/resource-centre/pdf/Influenza_A_Cleavage_Sites.pdf> (2018).

13 Stern, D. L. The genetic causes of convergent evolution. *Nat Rev Genet* **14**, 751-764, doi:10.1038/nrg3483 (2013).

14 Gutierrez, B., Escalera-Zamudio, M. & Pybus, O. G. Parallel molecular evolution and adaptation in viruses. *Current Opinion in Virology* **34**, 90-96, doi:10.1016/j.coviro.2018.12.006. (2019).

15 Pond, S. L., Murrell, B. & Poon, A. F. Evolution of viral genomes: interplay between selection, recombination, and other forces. *Methods Mol Biol* **856**, 239-272, doi:10.1007/978-1-61779-585-5_10 (2012).

16 Frost, S. D. W., Magalis, B. R. & Kosakovsky Pond, S. L. Neutral Theory and Rapidly Evolving Viral Pathogens. *Mol Biol Evol* **35**, 1348-1354, doi:10.1093/molbev/msy088 (2018).

17 Bloom, J. D., Gong, L. I. & Baltimore, D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* **328**, 1272-1275, doi:10.1126/science.1187816 (2010).

18 Nao, N. *et al.* Genetic Predisposition To Acquire a Polybasic Cleavage Site for Highly Pathogenic Avian Influenza Virus Hemagglutinin. *MBio* **8**, doi:10.1128/mBio.02298-16 (2017).

19 Bhattacharya, T. *et al.* Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* **315**, 1583-1586, doi:10.1126/science.1131528 (2007).

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

21 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).

22 Martin, D. P. *et al.* RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462-2463, doi:10.1093/bioinformatics/btq467 (2010).

23 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, doi:10.1093/bioinformatics/bty633 (2018).
24 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**, 1969-1973, doi:10.1093/molbev/mss075 (2012).
25 Ferreira, M. A. R. & Suchard, M. A. Bayesian analysis of elapsed times in continuous-time Markov chains. *Canadian Journal of Statistics* **36**, 355-368, doi:https://dx.doi.org/10.1002/cjs.5550360302 (2008).
26 Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS Comput Biol* **5**, e1000520, doi:10.1371/journal.pcbi.1000520 (2009).
27 Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology*. (Oxford University Press, 1991).
28 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591, doi:10.1093/molbev/msm088 (2007).
29 Delport, W., Poon, A. F., Frost, S. D. & Kosakovsky Pond, S. L. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**, 2455-2457, doi:10.1093/bioinformatics/btq429 (2010).
30 Kosakovsky Pond, S. L. & Frost, S. D. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* **22**, 1208-1222, doi:10.1093/molbev/msi105 (2005).
31 Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* **8**, e1002764, doi:10.1371/journal.pgen.1002764 (2012).
32 Murrell, B. *et al.* Gene-wide identification of episodic selection. *Mol Biol Evol* **32**, 1365-1371, doi:10.1093/molbev/msv035 (2015).
33 Smith, M. D. *et al.* Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol* **32**, 1342-1353, doi:10.1093/molbev/msv022 (2015).
34 Kosakovsky Pond, S. L., Poon, A. F., Leigh Brown, A. J. & Frost, S. D. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol Biol Evol* **25**, 1809-1824, doi:10.1093/molbev/msn123 (2008).
35 Reich, S. *et al.* Structural insight into cap-snatching and RNA synthesis by influenza polymerase. *Nature* **516**, 361-+, doi:10.1038/nature14009 (2014).
36 Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat Methods* **12**, 7-8, doi:10.1038/nmeth.3213 (2015).
37 Bloom, J. D. & Glassman, M. J. Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. *PLoS Comput Biol* **5**, e1000349, doi:10.1371/journal.pcbi.1000349 (2009).
38 Xu, Y. *et al.* Limited Antigenic Diversity in Contemporary H7 Avian-Origin Influenza A Viruses from North America. *Sci Rep* **6**, 20688, doi:10.1038/srep20688 (2016).
39 Bosch, F. X., Garten, W., Klenk, H. D. & Rott, R. 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 (1981).
40 Nguyen, T. *et al.* Characterization of a highly pathogenic avian influenza H5N1 virus sublineage in poultry seized at ports of entry into Vietnam. *Virology* **387**, 250-256, doi:10.1016/j.virol.2009.03.006 (2009).
41 Zhao, G. *et al.* Novel reassortant highly pathogenic H5N2 avian influenza viruses in poultry in China. *PLoS One* **7**, e46183, doi:10.1371/journal.pone.0046183 (2012).
42 Uchida, Y. *et al.* Genetic characterization and susceptibility on poultry and mammal of H7N6 subtype avian influenza virus isolated in Japan in 2009. *Vet Microbiol* **147**, 1-10, doi:10.1016/j.vetmic.2010.05.037 (2011).
43 Monne, I. *et al.* Emergence of a highly pathogenic avian influenza virus from a low-pathogenic progenitor. *J Virol* **88**, 4375-4388, doi:10.1128/JVI.03181-13 (2014).
44 Anisimova, M. & Kosiol, C. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol* **26**, 255-271, doi:10.1093/molbev/msn232 (2009).
45 Linster, M. *et al.* Identification, characterization, and natural selection of mutations driving airborne transmission of A/H5N1 virus. *Cell* **157**, 329-339, doi:10.1016/j.cell.2014.02.040 (2014).
46 Saenz, R. A. *et al.* Quantifying transmission of highly pathogenic and low pathogenicity H7N1 avian influenza in turkeys. *PLoS One* **7**, e45059, doi:10.1371/journal.pone.0045059 (2012).

47 Long, J. S. *et al.* The effect of the PB2 mutation 627K on highly pathogenic H5N1 avian influenza virus is dependent on the virus lineage. *J Virol* **87**, 9983-9996, doi:10.1128/JVI.01399-13 (2013).
48 Yang, H., Chen, L. M., Carney, P. J., Donis, R. O. & Stevens, J. Structures of receptor complexes of a North American H7N2 influenza hemagglutinin with a loop deletion in the receptor binding site. *PLoS Pathog* **6**, e1001081, doi:10.1371/journal.ppat.1001081 (2010).
49 Lu., J. *et al.* Molecular evolution, diversity and adaptation of H7N9 influenza A viruses in China. *Emerging Infectious Diseases* **24**, doi:DOI: 10.3201/eid2410.171063 (2018).
50 Ha, Y., Stevens, D. J., Skehel, J. J. & Wiley, D. C. H5 avian and H9 swine influenza virus haemagglutinin structures: possible origin of influenza subtypes. *EMBO J* **21**, 865-875, doi:10.1093/emboj/21.5.865 (2002).
51 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).
52 Katz, J. M. *et al.* Molecular correlates of influenza A H5N1 virus pathogenesis in mice. *J Virol* **74**, 10807-10810 (2000).
53 Melen, K. *et al.* Nuclear and nucleolar targeting of influenza A virus NS1 protein: striking differences between different virus subtypes. *J Virol* **81**, 5995-6006, doi:10.1128/JVI.01714-06 (2007).
54 Stevens, J. *et al.* Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. *Science* **312**, 404-410, doi:10.1126/science.1124513 (2006).
55 Hoffmann, E., Lipatov, A. S., Webby, R. J., Govorkova, E. A. & Webster, R. G. Role of specific hemagglutinin amino acids in the immunogenicity and protection of H5N1 influenza virus vaccines. *Proc Natl Acad Sci U S A* **102**, 12915-12920, doi:10.1073/pnas.0506416102 (2005).
56 Saleh, M. M., Selim, A., Shahera, A. A., El -Shahidy, M. S. & El-Trabilli, M. M. Molecular characterization of recently isolated highly pathogenic avian influenza virus (H5N1) in Egypt. *Nature and Science* **15** (2017).
57 Perdue, M. L. & Suarez, D. L. Structural features of the avian influenza virus hemagglutinin that influence virulence. *Vet Microbiol* **74**, 77-86 (2000).
58 Steel, J., Lowen, A. C., Mubareka, S. & Palese, P. Transmission of influenza virus in a mammalian host is increased by PB2 amino acids 627K or 627E/701N. *PLoS Pathog* **5**, e1000252, doi:10.1371/journal.ppat.1000252 (2009).
59 Nilsson, B. E., Te Velthuis, A. J. & Fodor, E. Role of the PB2 627 Domain in Influenza A Virus Polymerase Function. *J Virol* **91**, doi:10.1128/JVI.02467-16 (2017).
60 Lee, C. Y. *et al.* Prerequisites for the acquisition of mammalian pathogenicity by influenza A virus with a prototypic avian PB2 gene. *Sci Rep* **7**, 10205, doi:10.1038/s41598-017-09560-z (2017).
61 Lebarbenchon, C. & Stallknecht, D. E. Host shifts and molecular evolution of H7 avian influenza virus hemagglutinin. *Virol J* **8**, 328, doi:10.1186/1743-422X-8-328 (2011).
62 Fusaro, A. *et al.* Evolutionary trajectories of two distinct avian influenza epidemics: Parallelisms and divergences. *Infect Genet Evol* **34**, 457-466, doi:10.1016/j.meegid.2015.05.020 (2015).
63 Banks, J. *et al.* 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 (2001).
64 Casadevall, A. & Imperiale, M. J. Risks and benefits of gain-of-function experiments with pathogens of pandemic potential, such as influenza virus: a call for a science-based discussion. *MBio* **5**, e01730-01714, doi:10.1128/mBio.01730-14 (2014).
65 Veits, J. *et al.* Avian influenza virus hemagglutinins H2, H4, H8, and H14 support a highly pathogenic phenotype. *Proc Natl Acad Sci U S A* **109**, 2579-2584, doi:10.1073/pnas.1109397109 (2012).
66 Hill, S. C. *et al.* Antibody responses to avian influenza viruses in wild birds broaden with age. *Proc Biol Sci* **283**, doi:10.1098/rspb.2016.2159 (2016).
67 Wikramaratna, P. S., Pybus, O. G. & Gupta, S. 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, doi:10.1073/pnas.1401849111 (2014).
68 Thompson, C. P. *et al.* A naturally protective epitope of limited variability as an influenza vaccine target. *Nat Commun* **9**, 3859, doi:10.1038/s41467-018-06228-8 (2018).

1 **Large-scale phylogenetic analysis:**
Large-scale dataset compilation and initial phylogenetic analyses

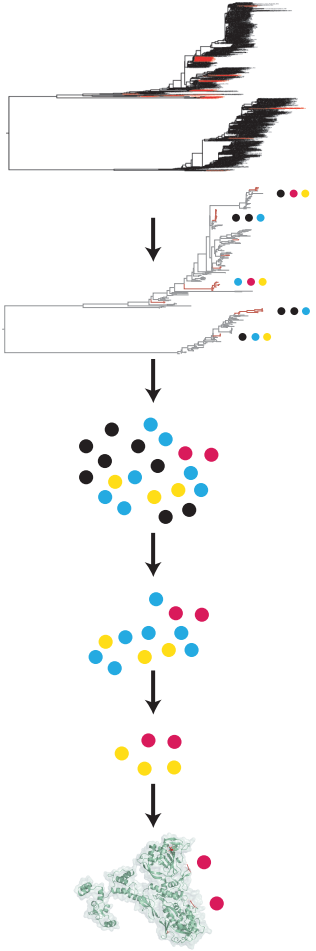
2 **Phylogenetically-informed subsmapling:**
to reduce tree size, yet retaining the mutational pathways of interest

3 **Detecting HAPMs occurring across HP clusters:**
using ancestral state reconstruction (ML)

4 **Testing for genotype-to-phenotype associations:**
under a model that considers the phylogenetic history of the trait and the phenotype

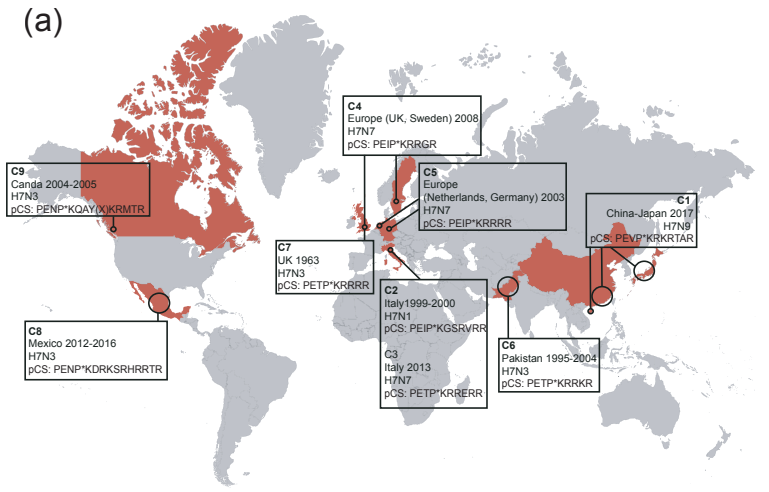
5 **Testing for diversifying positive selection:**
selecting parallel genetic changes scored as PSS

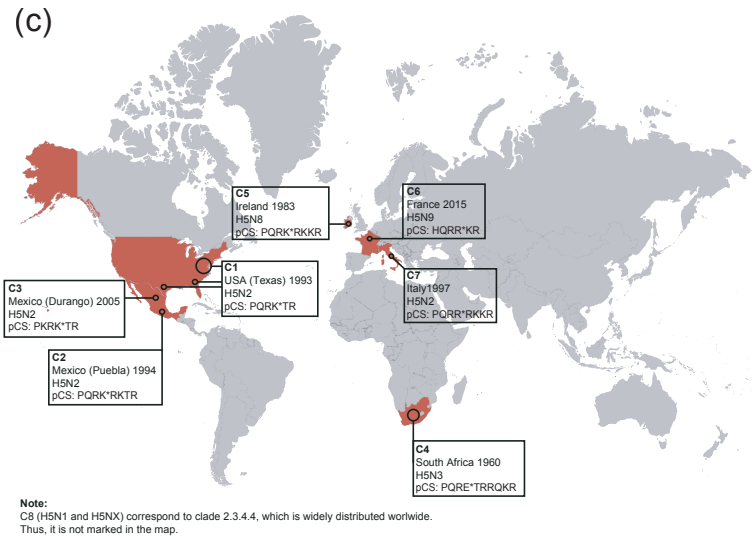
6 **Inferring mutation effects on protein stability and function:**
using a structural biology approach

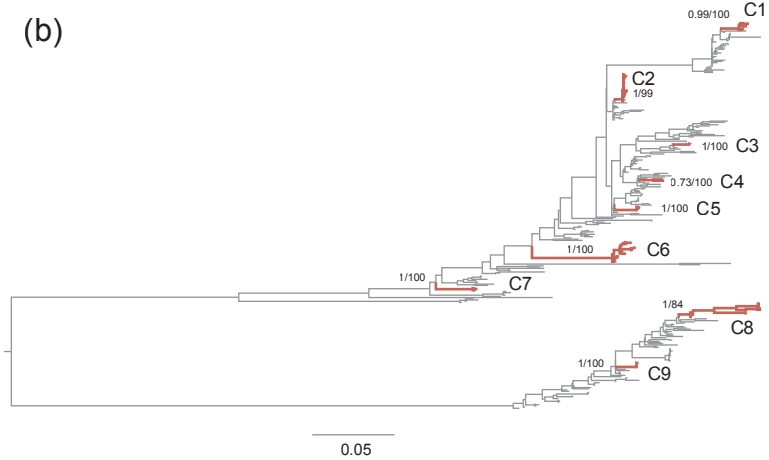


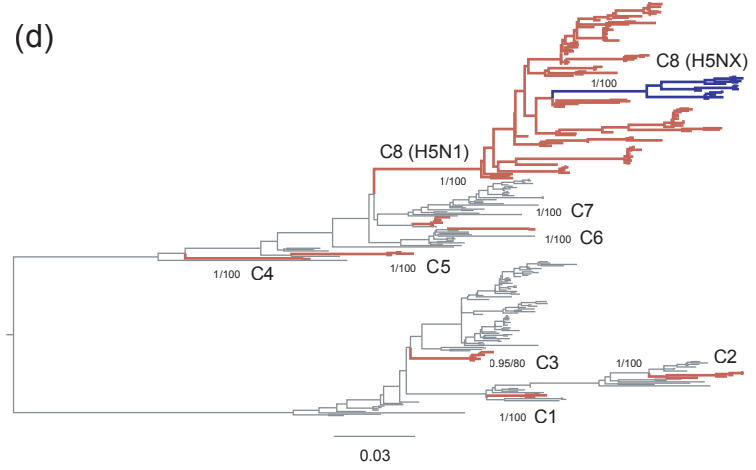
H7NX

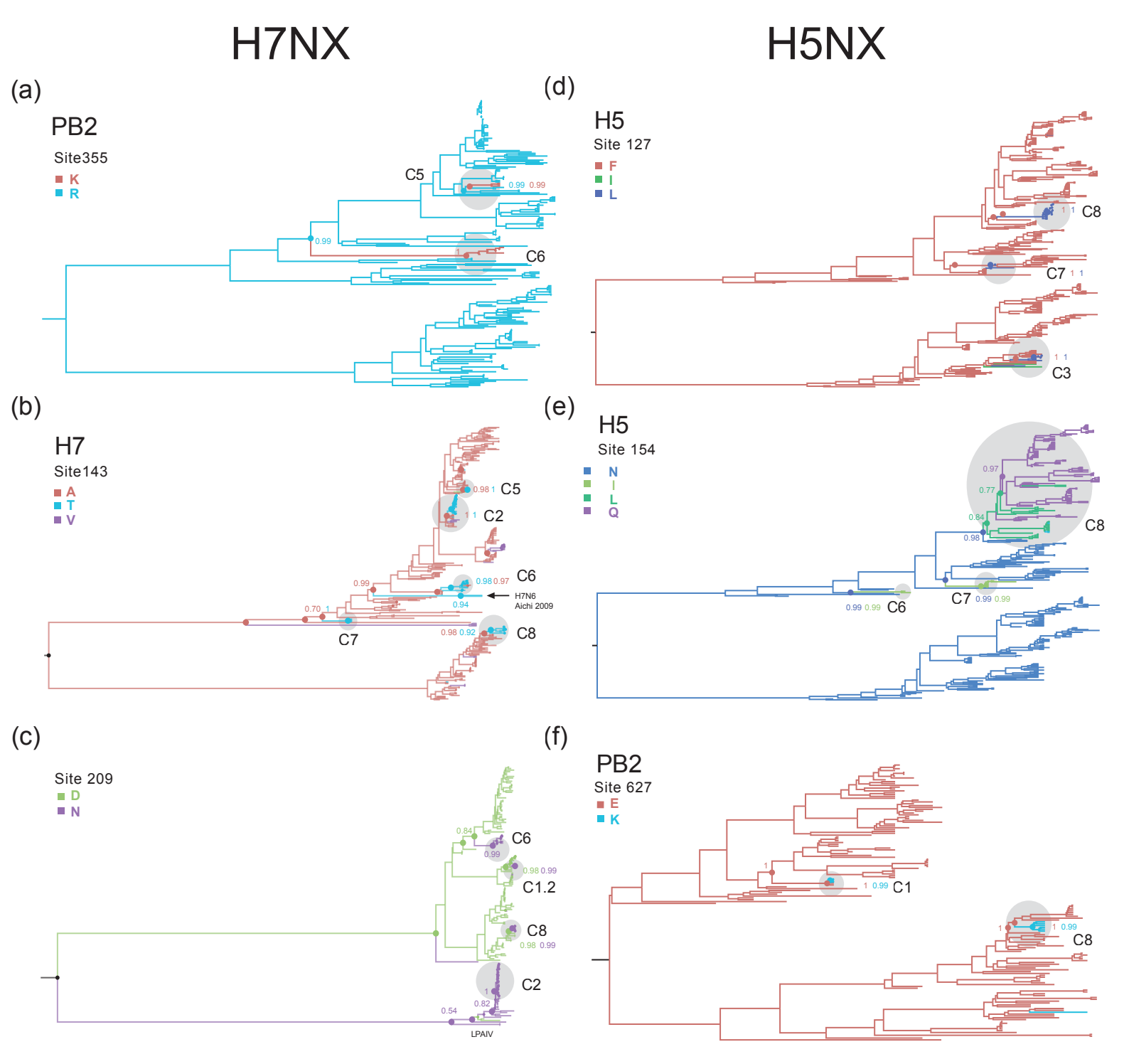
H5NX

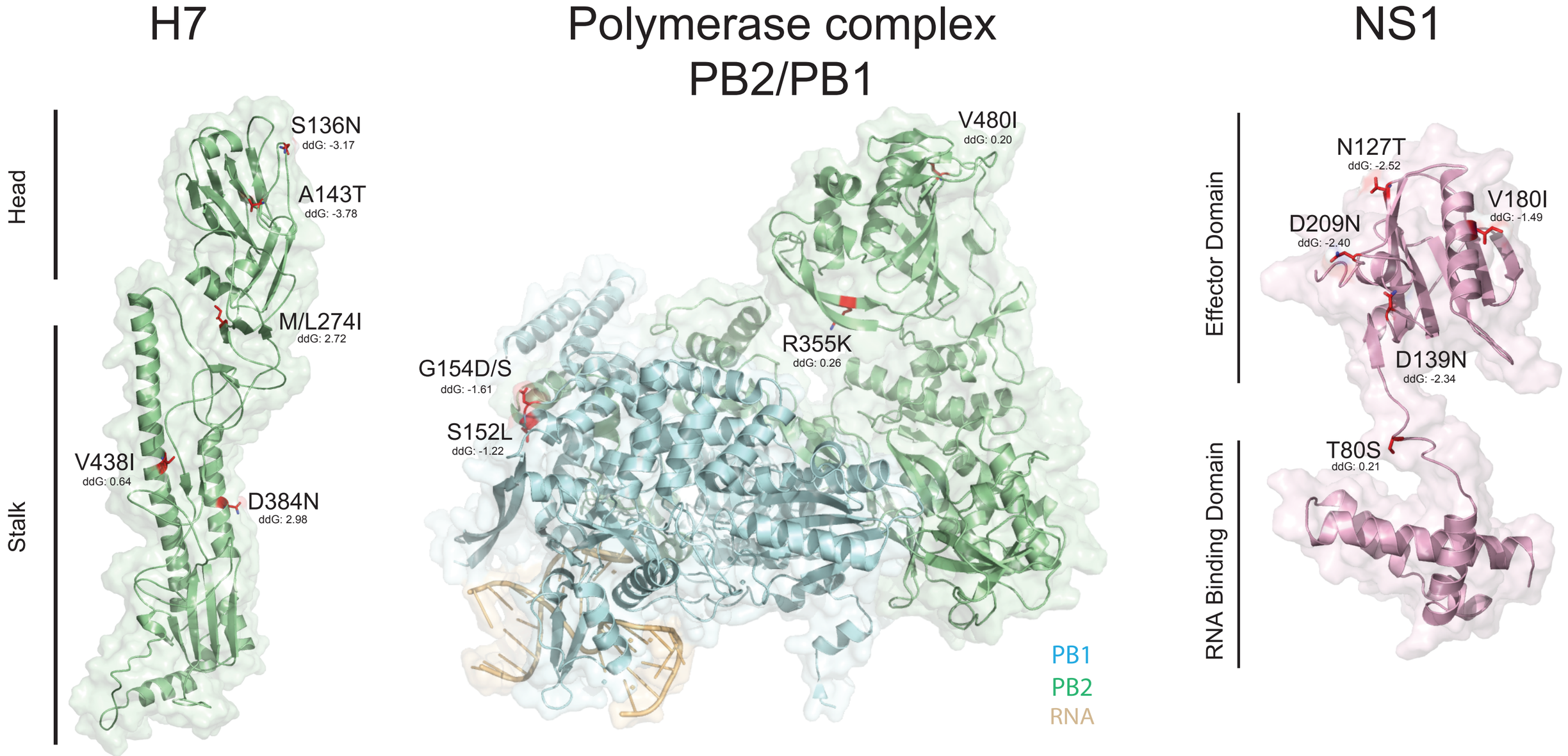

(a)
C9
Canda 2004-2005
H7N3
pCS: PENP*KQAY(X)KRMTR
C4
Europe (UK, Sweden) 2008
H7N7
pCS: PEIP*KRRGR
C5
Europe
(Netherlands, Germany) 2003
H7N7
pCS: PEIP*KRRRR
C1
China-Japan 2017
H7N9
pCS: PEVP*KRKRTAR
C7
UK 1963
H7N3
pCS: PETP*KRRRR
C2
Italy1999-2000
H7N1
pCS: PEIP*KGSRVRR
C3
Italy 2013
H7N7
pCS: PETP*KRRERR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


(c)
C5
Ireland 1983
H5N8
pCS: PQRK*RKKR
C6
France 2015
H5N9
pCS: HQRR*KR
C1
USA (Texas) 1993
H5N2
pCS: PQRK*TR
C3
Mexico (Durango) 2005
H5N2
pCS: PKRK*TR
C7
Italy1997
H5N2
pCS: PQRR*RKKR
C2
Mexico (Puebla) 1994
H5N2
pCS: PQRK*RKTR
C4
South Africa 1960
H5N3
pCS: PQRE*TRRQKR
Note:
C8 (H5N1 and H5NX) correspond to clade 2.3.4.4, which is widely distributed worlwide.
Thus, it is not marked in the map.


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


(d)
C8 (H5NX)
1/100
C8 (H5N1)
1/100
1/100
C7
1/100
C6
1/100
C4
1/100
C5
0.95/80
C3
1/100
C2
1/100
C1
0.03


H7NX
H5NX
(a)
PB2
Site355
K
R
C5
C6
(b)
H7
Site143
A
T
V
C5
C2
C6
H7N6
Aichi 2009
C7
C8
(c)
Site 209
D
N
C6
C1.2
C8
C2
LPAIV
(d)
H5
Site 127
F
I
L
C8
C7
C3
(e)
H5
Site 154
N
I
L
Q
C8
C7
C6
(f)
PB2
Site 627
E
K
C1
C8


H7
Polymerase complex
PB2/PB1
NS1
Head
Stalk
S136N
ddG: -3.17
A143T
ddG: -3.78
M/L274I
ddG: 2.72
V438I
ddG: 0.64
D384N
ddG: 2.98
V480I
ddG: 0.20
R355K
ddG: 0.26
G154D/S
ddG: -1.61
S152L
ddG: -1.22
PB1
PB2
RNA
Effector Domain
RNA Binding Domain
N127T
ddG: -2.52
V180I
ddG: -1.49
D209N
ddG: -2.40
D139N
ddG: -2.34
T80S
ddG: 0.21

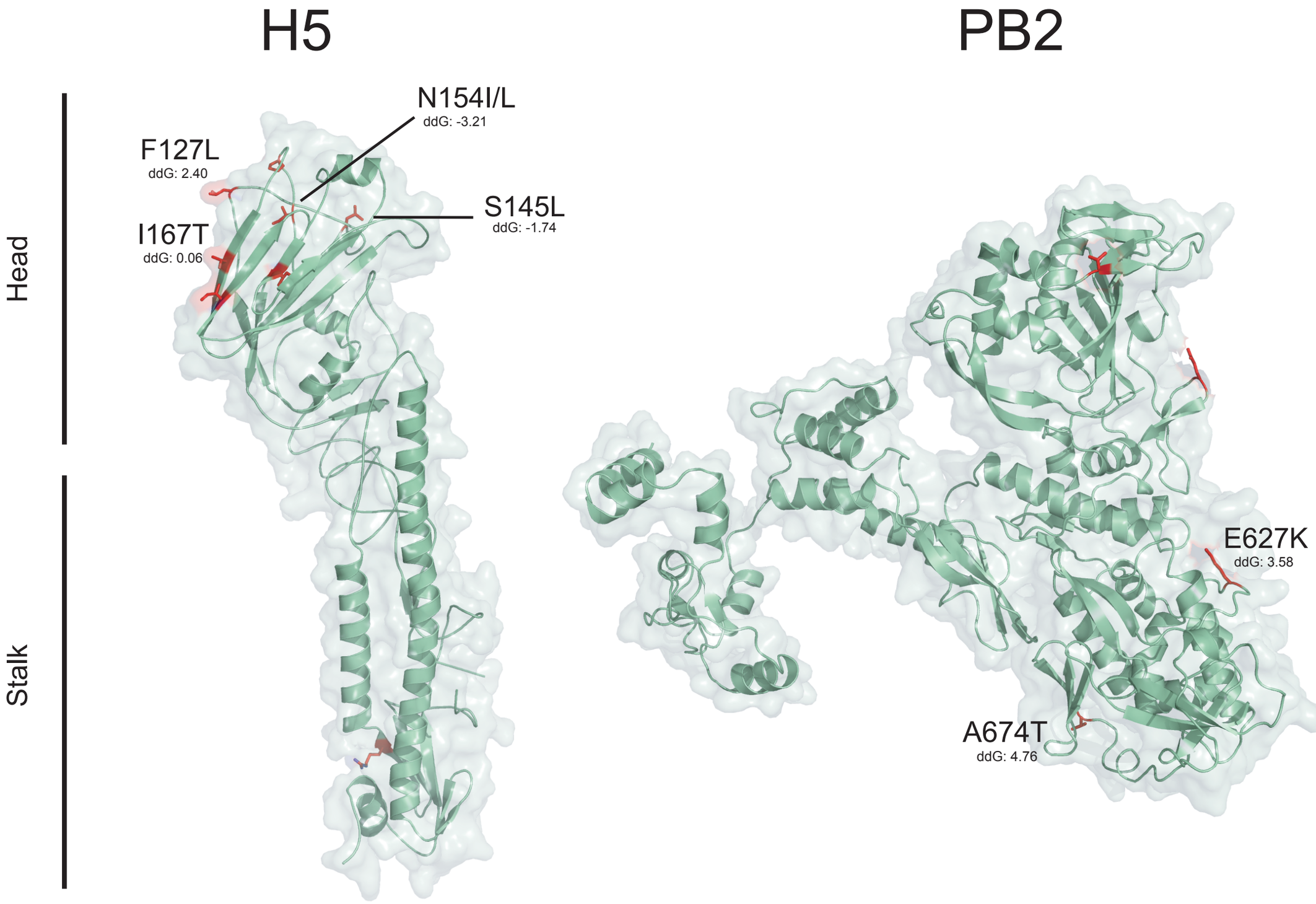

H5
PB2
Head
Stalk
N154I/L
ddG: -3.21
F127L
ddG: 2.40
S145L
ddG: -1.74
I167T
ddG: 0.06
E627K
ddG: 3.58
A674T
ddG: 4.76

TABLE 1. HAPMs significantly associated with the HP phenotype of H5 and H7 virus subtypes

H7NX							
HAPM	**Segment**	**HP clusters**	**Num. HP clusters**	**λ [95% CI]**	**p-value**	**H7 numbering**	**Reference***
A143T †	HA	C2, C5, C6, C7 and C8	5	7.1 [3.3, 13.2]	0.00	125	67,68,69,75
D209N †	NS1	C1.1, C6 and C8	3	2.4 [1.0, 4.3]	0.03		
R355K †	PB2	C5 and C6	2	6.0 [2.8, 11.8]	0.00		57
M/L274I	HA	C4 and C8	2	6.2 [3.1, 11.5]	0.00	256	
V480I	PB2	C5 and C6	2	9.2 [4.3, 18.9]	0.00		
V438I	HA	C3 and C8	2	4.6 [2.4, 8.3]	0.01	423	
D384N	HA	C6 and C8	2	5.1 [2.6, 9.5]	0.02	369	
G154D	PB1	C2 and C6	2	7.7 [4.3, 13.2]	0.03		67
S152L	PB1	C6 and C9	2	3.1 [1.7, 5.2]	0.03		
H5NX							
HAPM	**Segment**	**HP clusters**	**Num. HP clusters**	**λ [95% CI]**	**p-value**	**H5 numbering**	**Reference***
V113I	PB1	C2, C4 and C8	3	7.3 [3.7, 14.5]	0.01		
F127L†	HA	C3, C7 and C8	3	6.5 [3.3, 12]	0.03	111	59
N154I/L/Q† §	HA	C6, C7 and C8	3	See footnote	0.02	138	59
S145L/del‡	HA	C5 and C8	2	See footnote	0.01	129	62,75
A674T	PB2	C2 and C6	2	7.4 [3.9, 13.6]	0.00		65
M242I	HA	C1 and C8	2	7.4 [3.7, 13.7]	0.00	226	72
I167T	HA	C6 and C8	2	5.8 [3.0, 10.4]	0.03	151	61
E627K	PB2	C1 and C8	2	5.1 [2.7, 9.2]	0.04		
I451T	PB2	C2 and C8	2	3.8 [2.1, 6.8]	0.03		
R101K	M	C5 and C8	2	3.7 [1.7, 6.9]	0.02		
T204I	HA	C2 and C8	2	2.7 [1.4, 4.7]	0.03	188	
K379R	HA	C1 and C8	2	14.8 [7.4, 27.5]	0.00	363	

†Scored as positively selected site under different methods (see Supplementary Table 1 in the Supplementary Information).

§ Individual lambda values for each site are as follow: for 154I 1.7 [1.0, 2.7], for 154L 3.2 [1.9, 5.2] and for 154Q: 3.4 [2.0, 5.8], all with a p-value of 0.00.

‡ Individual lambda values for each site are as follow: for 145L 77.5 [69.5, 100.0] with a p-value of 0.00, and for 145del 7.0 [0.5, 14.7] with a p-value of 0.04.

* Where such HAPMs had been previously mentioned.