**Evolution of H5 highly pathogenic avian influenza: sequence data indicate stepwise changes in the cleavage site**

**Celia Abolnik**

Department of Production Animal Studies, Faculty of Veterinary Science, University of Pretoria, Onderstepoort, 0110, South Africa

Postal address: Faculty of Veterinary Science, Private Bag X04, Onderstepoort, 0110, Pretoria

Email: celia.abolnik@up.ac.za

Tel: +27 125298258

Fax: +27 125298306

**Abstract**

The genetic composition of an H5 subtype hemagglutinin gene quasispecies, obtained from ostrich tissues that had been infected with H5 subtype influenza virus was analysed using a next generation sequencing approach. The first evidence for the reiterative copying of a poly (U) stretch in the connecting peptide region in the haemagglutinin cleavage site (HACS) by the viral RNA-dependent RNA polymerase is provided. Multiple non-consensus species of RNA were detected in the infected host, corresponding to likely intermediate sequences between the putative low pathogenic precursor nucleotide sequence of the H5 influenza strain and the highly pathogenic avian influenza virus gene sequence. *In silico* analysis of the identified RNA sequences predicted that the intermediary H5 sequence PQREKRGLF plays an important role in subsequent mutational events that relocate the HACS coding region from stable base-paired RNA regions to a single-stranded bulge, thereby priming the connecting peptide coding region for RdRp slippage.

## 1. Introduction

The influenza A virus haemagglutinin protein (HA) is an integral membrane protein that forms spiked projections on the viral particle. Its main role in the viral life cycle is to attach to host cell receptors and facilitate fusion of the viral envelope and endosomal membrane to initiate the infection cycle [1]. Following synthesis as a single polypeptide chain, designated HA0, the protein undergoes various posttranslational modifications (e.g. glycosylation; palmitoylation) the last of which is proteolytic cleavage of HA0 into two disulphide-linked subunits, HA1 and HA2.The cleavage exposes the free amino terminus of HA2, a structure critical to virus-cell fusion and therefore infectivity [2,3].

The hemagglutin cleavage site (HACS) sequence in HA0 is cleaved by host proteases. In low pathogenicity (LPAI) subtypes of influenza A virus (IAV), the HACS motif typically contains a single arginine residue, i.e. Q-R/K-X-T-R (where X is a nonbasic amino acid). This motif is recognised by trypsin-like proteases that are mainly expressed in the respiratory and intestinal tracts, and therefore LPAI viruses produce localized infections with asymptomatic or mild effects. Highly pathogenic forms of the virus (HPAI) contain multibasic HACS sequences, i.e. Q-R/K-X-R/K-R [4]. Although sixteen subtypes of IAV are known to infect avian species (H1 to H16) [5], the HACS of only the H5 and H7 subtypes are prone to acquire multibasic motifs, derived from the exchange and insertion of basic amino acids. This produces a major shift in the pathogenic potential of the virus, as the altered HACS motif is recognised and cleaved by an alternative subset of proteases in the host. Multibasic motifs are recognised by subtilisin-related endoproteases that expressed in the Golgi and/or trans-Golgi network and members of this family include furin, PC6, mosaic serine protease large (MSPL) and transmembrane protease 13 (TMPRSS13) [6]. The broad tissue expression of these proteases enables systemic viral replication and consequently the highly infectious and lethal disease referred to as highly pathogenic avian influenza (HPAI) [3,7]. Although the pathogenicity of IAV is ultimately a polygenic trait, the HACS motif remains a prime determinant [8].

Once LPAI H5 and H7 strains are introduced from their natural aquatic bird reservoir into susceptible terrestrial avian hosts, the HPAI forms emerge spontaneously after a period of circulation within the flock [9]. Such events are described across the world in chickens, turkeys and ostrich flocks [10-15] and evidence for the conversion to HPAI from LPAI by the incorporation of multi-basic motifs into the HACS in field strains is supported both by *in vitro* and *in vivo* clinical studies [7, 16]. In rare cases field cases non-homologous recombination resulting in the insertion of a foreign nucleotide sequence into the HACs have been reported, with donor sequences derived from host 28S RNA [17], the viral matrix protein [18] or the viral nucleoprotein gene [19]. Notwithstanding non-homologous recombination, the mechanism of how multiple insertions of basic amino acids occur and why insertions are restricted to the HACs of H5 and H7 strains remains obscure. Two main theories have been put forward: (a) purine triplets that are duplications of existing sequences are incorporated into the HACs during strand slippage of the polymerase complex during transcription [11, 20] and (b) basic amino acids are progressively accumulated by a stepwise process involving amino acid substitutions [21, 22]. From these studies there appears to be a general consensus that RNA secondary structure and polymerase slippage are at play in the generation of multi-basic HACSs, albeit in the absence of empirical evidence.

In 2011, an HPAI H5N2 outbreak affected farmed ostriches in South Africa's Western Cape Province. The index case was an ostrich chick that died from the infection on the 3$^{rd}$ of March 2011, and organ samples from this bird had been subjected to Next Generation Sequencing [13]. In this paper the quasispecies population at the HACS was analysed and *in silico* analyses of the identified RNA sequences were performed in order to form a basis for the underlying mechanism supporting the emergence of H5 HPAI H5 from a putative LPAI progenitor.

## 2. Materials and Methods

### *2.1 Library preparation and Illumina sequencing*

The original homogenate of the ostrich tissue pool comprising trachea, heart, lung, liver, spleen and kidney collected on the 3rd of March 2011 was analysed. Total RNA was extracted using TriZOL® reagent (Invitrogen) and the transcriptome was amplified from 200ng of RNA as described previously [13]. Illumina sequencing was performed by the sequencing service provider, ARC-Biotechnology Platform, Pretoria. Briefly, the Illumina library was prepared from 55ng of cDNA using the Nextera sample preparation kit (Epicentre Biotechnologies). The libraries were purified using the QIAquick PCR purification kit (Qiagen) and quantified using a Qubit 2.0 fluorometer (Invitrogen). One lane was sequenced on an Illumina HiScanSQ system with V2 sequencing reagents and a mixture of Illumina and Nextera sequencing primers to produce paired-end reads with an average length of 97 nucleotides (nt).

### *2.2 Sequence analysis*

The CLC Genomics Workbench v7.5.2 was used for downstream bioinformatics analyses. Reads with an average length of 100nt were generated and after quality trimming, these were assembled against the reference sequence, segment 4 (HA) of isolate A/ostrich/South Africa/2114/2011, accession number JX069081. The HACS in the reference sequence spanned nucleotides 1037 to 1063, therefore mapped reads spanning position 1000-1080 were extracted and imported into BioEdit [23]. Reads were manually inspected, reverse-complemented where necessary and aligned. Only reads that spanned the entire cleavage site i.e. beginning with the CCU proline codon and ending with the UUU phenylalanine codon were considered for this analysis. The reason for not using partial sequences at the HACS was so that haplotypes, i.e. mutations that are inherited together, could be assessed as this would affect RNA folding predictions. Viral complementary RNA (cRNA) sequences were converted into the complement viral genomic sequence (vRNA) for RNA folding in BioEdit. Sequence JX069081, which corresponds to the master sequence in the quasipecies

distribution was modified *in silico* with the addition of standard complete 3' and 5' terminal sequences [24]. The hypothetical LPAI progenitor was constructed by substituting the HACS with base pairs encoding default LPAI sequence PQRETRGLF. RNA structure was predicted the CLC Genomics Workbench v7.5.2 that uses a two-step algorithm [25]. A minimum free energy approach without base pairing constraints was applied.

## 3. Results

### *3.1 Quasispecies at the HACS*

Next generation sequencing was applied to tissue extracts from the original infected ostrich in the 2011 outbreak index case. 1,319,369 reads were produced with an average length of 97nt, and 117,279 (8.89%) of these mapped to the complete HA gene sequence. A subset of 340 reads spanning the entire HACS sequence was retrieved. The complete list of HACS cDNA nucleotide sequences arranged by length, the translated amino acid motif and frequency of each variant is presented in Table 1. Notably no HACS motif for the LPAI progenitor (PQRETRGLF) or any other HACS sequences of 27 nucleotides were detected. Similarly, during an assessment of the quasispecies present in viral populations of the Italian HPAI H7N1 outbreak in 1999/2000, no evidence of the LPAI H7 progenitor sequence was detected in the HPAI samples [14]. Collectively these findings suggest that the LPAI sequence is under strong negative selection pressure soon after HPAI emerges in the quasispecies population, but this warrants further investigation.

The mutant distribution in the ostrich tissue ranged from 28nt to 39nt, the latter encoding the longest HACS motif of PQRRKKKKKKGLF. The region between the glutamine at -5 and glycine at +1 is referred to as the connecting peptide, thus the connecting peptide of the longest viable HACS is eight amino acids in length, whereas the connecting peptide of the master sequence, PQRRKKRGLF, is five amino acids in length. This master sequence of 30nt was present in the ostrich tissue at a frequency of 62%. Point mutations occurred at virtually every position in the region analysed (Table 1). As is typical of RNA viruses, the RNA-dependent RNA polymerase (RdRp) of IAV is error-prone and

**Table 1.** H5 subtype influenza A virus quasispecies at the HACS of pooled ostrich tissues

| No.[a] | cDNA nucleotide sequence[b] | Amino acid sequence[c] | Number (frequency) | No.[a] | cDNA nucleotide sequence | Amino acid sequence | Number (frequency) |
|---|---|---|---|---|---|---|---|
| **27** | CCTCAAAGA-G----------AAACAAGAGGTCTATTT[d] | PQRETRGLF | 0 | 30/… | ACTCCAAGAAG---------AAAAAAAAGAGGTCTATTT | TPRRKKRGLF | 1 (0.29) |
| 28 | CCTCAAAGAAG----------AAAAAAGAGGTCTATTT | PQRRKKRSI | 3 (0.88) | | CCTCAAAGAAG---------AAAAAAAAGAGGTCAATTT | PQRRKKRGQF | 1 (0.29) |
| 29 | CCTCAAAGAAG----------AAAATAAGAGGTCTATTT | PQRRK*EVY | 1 (0.29) | | CATCAAAGAAG---------AAAAAAAAGAGGTCTATTT | **HQRRKKRGLF** | 2 (0.58) |
| | CCTCAAAGAAG----------AAAAAAAAAGGTCTATTT | PQRRKKKVY | 1 (0.29) | | ACTCAAAGAAG---------AAAAAAAAGAGGTCTATTT | **TQRRKKRGLF** | 1 (0.29) |
| | CCTCAAAGAAG----------AAAAAAAGAGGTCTATTT | PQRRKKEVY | 23 (6.76) | | CCTAAAAAAAA---------AAAAAAAAGAGGTCTATTT | PKKKKKRGLF | 1 (0.29) |
| | CCTCAAAGAAG----------AAAAAAAGAGGACTATTT | PQRRKKEDY | 1 (0.29) | | CCTCAAAGAAG---------AAAAAAATGAGATCTATTT | PQRRKK*DLF | 1 (0.29) |
| 30 [5] | CCTCAAAGAAG---------AAAAAAAGGAGGTCTATTT | PQRRKKGGLF | 1 (0.29) | | CCTCAAAGAAG---------AAAACAAAGAGGTCTATTT | **PQRRKQRGLF** | 1 (0.29) |
| | CCTCAAAGAAG---------AAAAAAAAGAGGGCTATTT | **PQRRKKRGLF** | 1 (0.29) | | CCTCAAAGAAG---------AAGAAAAAGAGGTCTATTT | **PQRRRKRGLF** | 2 (0.58) |
| | CCTCAAAGAAG---------AAAAAAAAGATGTCTATTT | PQRRKKRCLF | 1 (0.29) | | CCTCAAAGAAT---------AAAAAAAAGAGGTCTATTT | **PQRIKKRGLF** | 1 (0.29) |
| | CCTCAAAGAAT---------AAAAAAAAGAGGTCTATTT | **PQRIKKRGLF** | 1 (0.29) | | CCTCAAAGCAA---------AAAAAAAAGAGGTCTATTT | **PQSKKKRGLF** | 1 (0.29) |
| | CCTCAAAGAAG---------AAAATAAAGAGGTCTATTT | PQRRK*RGLF | 1 (0.29) | | CCTCAAAGAAG---------AACAAAAAGAGGTCTATTT | **PQRRTKRGLF** | 1 (0.29) |
| | CCTCAAAGAAG---------AAAAAAAAGAGGTCTATCT | PQRRKKRGLS | 1 (0.29) | | CCTCAAAGAAG---------AAAATAAAGAGGTCTATTT | PQRRK*RGLF | 1 (0.29) |
| | CTTCAAAGAAG---------AAAAAAAAGAGGTCTATTT | **LQRRKKRGLF** | 1 (0.29) | | CCTCAAAGAAG---------AAAAATAAGAGGTCTATTT | **PQRRKIRGLF** | 1 (0.29) |
| | CCTCAATGAAG---------AAAAAAAAGAGGTCTATTT | PQ*RKKRGLF | 1 (0.29) | | CCTCAAAGAAG---------AAAAAATAGAGGTCTATTT | **PQRRKNRGLF** | 1 (0.29) |
| | CCCCAATGAAG---------AAAAAAAAGAGGTCTATTT | PQ*RKKRGLF | 1 (0.29) | | CCTCAAGGAAG---------AAAAAAAAGAGGTCTATTT | **PQGRKKRGLF** | 1 (0.29) |
| | CCTCAATGAAG---------AAATAAAAGAGGTCTATTT | PQ*RNKRGLF | 1 (0.29) | | CCTCAAAGAAG---------AAAAAGAAGAGGTCTATTT | **PQRRKRRGLF** | 1 (0.29) |
| | CCTCATAGAAG---------AAAAAAAAGAGGTCTATTT | PHRRKKRGLF | 1 (0.29) | | CCTCAAAGAAG---------AAAAAAAAGAGGTTTATTT | **PQRRKKRGLF** | 1 (0.29) |
| | CCTCAAAGAAG---------AAAAAAAAGATGTCTATTT | PQRRKKRCLF | 1 (0.29) | | CCTCAAAGAAG---------AAAAAAGAGAGGTCTATTT | **PQRRKKRGLF** | 1 (0.29) |
| | CCTCAAAGAAG---------TAAAAAAAGTGGTCTATTT | PQRSKKSGLF | 1 (0.29) | | CCTCAGAGAAG---------AAAAAAAAGAGGTCTATTT | **PQRRKKRGLF** | 1 (0.29) |
| | TCTCAAAGAAG---------AAAAAAAAGAGGTCTATTT | SQRRKKRGLF | 1 (0.29) | | CCTCAAAGACG---------AAAAAAAAGAGGTCTATTT | **PQRRKKRGLF** | 1 (0.29) |
| | CCTCAAAGAAG---------AAAAGAAAGAGGTCTATTT | **PQRRKERGLF** | 2 (0.59) | | CCTCAAAGAAG---------AAAAAAAAGAGGCCTATTT | **PQRRKKRGLF** | 2 (0.58) |
| | CCTCAAAGAAG---------AAAAAAAAGAGGTCTATTA | PQRRKKRGLL | 1 (0.29) | | CCTCAAAGAAG---------AAAAAAACGAGGTCTATTT | **PQRRKKRGLF** | 1 (0.29) |
| | CCTCAAAGAAG---------AGAAAAAAGAGGTCTATTT | **PQRREKRGLF** | 3 (0.88) | | CCTCAAAGAAG---------AAAGAAAAGAGGTCTATTT | **PQRRKKRGLF** | 2 (0.58) |
| | CCTCAAAGAAG---------CAAAAAAAGAGGTCTATTT | **PQRSKKRGLF** | 1 (0.29) | | CCTCAAAGAAG---------AAAAAAAAGGGGACTATTT | **PQRRKKRGLF** | 1 (0.29) |
| | CCTCACAGACG---------AAAACAAAGAGGTCTATTT | PHRRKQRGLF | 1 (0.29) | | CCTCAAAGAAG---------AAAAAAAAGAGGTCTATTT | **PQRRKKRGLF** | 211 (62.05) |
| | CCTCAAAGAAG---------AAAACAAAGAGGTCTATTT | **PQRRKQRGLF** | 1 (0.29) | | AATCAAAGAAG---------AAAAAAAAGAGGTCTAATT | NQRRKKRGLI | 1 (0.29) |
| | CCTCACAGAAG---------AAAAAAAAGAGGTCTATTT | PHRRKKRGLF | 3 (0.88) | 31 | CCTCAAAGAAG--------AGAAAAAAAAGAGGTCTATT | PQRREKKEVY | 1 (0.29) |
| | CCTCAAAGAAG---------AAAAATAAGAGGTCTATTT | **PQRRKIRGLF** | 1 (0.29) | | CCTCAATGAAG--------AGAAAAAAAGAGGTCTATTT | PQ*REKKRSI | 1 (0.29) |
| | CCTCAAAGAAG---------AAATAAAAGAGGTCTATTT | **PQRRNKRGLF** | 1 (0.29) | | CCTCAAAGGAG--------AACAACAAAGAGGTCTATTT | PQRRTTKRSI | 1 (0.29) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | CCTCAAAGAAG---------AACAATAAGAGGTCTATTT | **PQRRTIRGLF** | 1 (0.29) | | CCTCAAAGAAG--------AAAAAAAAAGAGGTCTATTT | PQRRKKKRSI | 10 (2.94) |
| | ACTCAAAGAAG---------AACAAAAAGAGGTCTATTT | **TQRRTKRGLF** | 1 (0.29) | 32 | CCTCAAAGAAG-------ACAAAAAAAAGAGGTCTATTT | PQRRQKKEVY | 1 (0.29) |
| | CCTCAAAGAAG---------AAAAAAAAGCAGTCTATTT | PQRRKKSSLF | 1 (0.29) | | CCTCAAAGAAG-------AATAAAAAAAGAGGTCTATTT | PQRRIKKEVY | 1 (0.29) |
| | CCTCTAAAAAG---------AAAAAAAAGAGGTCTATTT | PLKRKKRGLF | 1 (0.29) | | CCTCAAAGAAG-------AAAAAAAAAAGAGGTCTATTT | PQRRKKKEVY | 2 (0.58) |
| | CCTCGAAGAAG---------AAAAAAAAGAGGTCTATTT | PRRRKKRGLF | 3 (0.88) | | CCTCAAAGAAG-------AGAAAAAAAAGAGGTCTATTT | PQRREKKEVY | 3 (0.88) |
| | ACTCGACGAAG---------AAAAAAAAGAGGTCTATTT | TRRRKKRGLF | 1 (0.29) | 33 [6] | CCTCAAAGAAG------AGAAAAAAAAAGAGGTCTATTT | **PQRREKKRGLF** | 1 (0.29) |
| | CCTCGAAGGAG---------ACAAAAAAGAGGTCTATTT | PRRRQKRGLF | 1 (0.29) | | CCTCAAAGAAG------AAAAAAAAAAAGAGGTCTATTT | **PQRRKKKRGLF** | 1 (0.29) |
| | CCTCAACGAAG---------ACATAAAAGAGGTCTATTT | **PQRRHKRGLF** | 1 (0.29) | 34 | CCTCAAAGAAG-----AAAAAAAAAAAAAGAGGTCTATTT | PQRRKKKKRSI | 1 (0.29) |
| | ACTCAAAGAAG---------AAAAAAAAGAGGTCTATTT | **TQRRKKRGLF** | 2 (0.58) | 35 | CCTCAAAGAAG----AAAAAAAAAAAAAGAGGCCTATTT | PQRRKKKKEAY | 1 (0.29) |
| | CCTCCAAGAAG---------AAAAAAAAGAGGTCTATTT | PPRRKKRGLF | 1 (0.29) | | CCTCAAAGAAG----AAAAAAAAAAAAAGAGGTCTATTT | PQRRKKKKEVY | 1 (0.29) |
| | CCTCAAAGCAG---------AAAAAAAAGAGGTCTATTT | **PQSRKKRGLF** | 1 (0.29) | 36 [7] | CCTCAAAGAAG---AAAAAAAAAAAAAAGAGGTCTATTT | **PQRRKKKKRGLF** | 1 (0.29) |
| | CCTCAAAGCAG---------AAAAAAAAGAGGTCTCTTT | **PQSRKKRGLF** | 1 (0.29) | 39 [8] | CCTCAAAGAAGAAAAAAAAAAAAAAAAAAAGGCTTATTT | **PQRRKKKKKKGLF** | 1 (0.29) |
| | CCTCAAAGTAG---------AAAAAAAAGAGGTCTATTT | **PQSRKKRGLF** | 1 (0.29) | | | | |

[a]Number of nucleotides in the HACS motif. Square brackets denote the number of amino acids in the connecting peptide.

[b]Point mutations are underlined

[c]Amino acid motifs that meet the criteria Q-R/K-X-R/K-R or Q-X-X-R-X-R/K-R or B(X)-X(B)-R/K-X-R/K-R, where B is a basic residue (Kido et al 2012), and produce an in-frame HA2 protein with the GLF motif at the amino terminus are highlighted. Where the amino acids motifs are identical, these are shaded. The nucleotides that are highlighted emphasise a clonal expansion of the inserted adenine residue between the two guanine residues. Nucleotide sequences are aligned to the longest 39-nt sequence, with (-) indicating gaps.

[d]The undetected hypothetical LPAI progenitor sequence

lacks proofreading function, leading to genome replication errors in the order of about 1 x $10^{-4}$ base substitutions per position per virus per generation, or about one base substitution in the HA gene per virus generation [26]. It cannot be excluded that a proportion of these mutations are artefacts of the transcriptome amplification or Illumina sequencing that escaped quality trimming, but generally the variation observed in this small region alone illustrates the remarkable stochastic adaptive potential of IAV.

### *3.2 The quasispecies at the HACS provides the first evidence of reiterative copying during replication*

The model proposed by Garcia et al [11] and Perdue et al [20] sought to explain how the AGA or AGG codons were introduced into the HACS during an H5 HPAI outbreak in Mexico in the 1990s. It was postulated that during replication three template adenines are copied into uracils, and these three uracils then base-pair with adenine in a downstream position. The inter-connecting sequence is thereby re-transcribed to generate the duplication of six bases that was observed in the HACS of Mexican viruses. The model restricts the insertion of nucleotides to multiples of three, but the current results reveal a spectrum of nucleotide insertions that are neither biased towards triplets nor are exact duplicates of an existing sequence in the HACS. Insertions ranged from a single nucleotide up to eleven, (with the exception that 9- and 10- nucleotide insertions weren't detected), pointing to an alternative mechanism that is more consistent with the progressive step-wise extension of the HACS during successive rounds of viral replication. It is not possible to determine whether in some cases more than one uracil residue was incorporated in a single event. Essentially, the insertion of uracils in the connecting peptide at the HACS more closely resembles the mechanism the IAV RdRp employs to polyadenylate its mRNAs: experimental studies established that as the RdRp nears the 5' end to which it is bound, it encounters steric hindrance from a conserved terminal hairpin loop structure adjacent to a uracil stretch on the viral RNA. The RdRp consequently stutters on the preceding stretch of uracils, which it repeatedly copies to produce a

poly (A) tail [27-30]. In the present study insertions of varying lengths were restricted to the uracil-rich connecting peptide region, thereby providing the first direct evidence of the reiterative copying of uracil residues in the HACS by the RdRp.

### *3.3 The heritability of HACSs containing non-sense or lethal mutations*

Following IAV entry and uncoating the vRNAs are transported into nucleus of the host cell to be transcribed into mRNA. vRNAs also serve as a template for cRNA, from which progeny vRNAs are produced. vRNAs subsequently either leave the nucleus to be incorporated into the budding virions [31]. In the original model for the conversion of LPAI to HPAI, the sequential incorporation of single nucleotides into the HACS was rejected on the basis that non-triplet nucleotide insertions wouldn't be viable in the population [20]. However, advances in the study of defective interfering (DI) particles allay this concern. DI particles are commonly described as virions that contain an internal deletion in at least one of their eight genome segments, as a consequence of erroneous translocation of the RdRp during transcription [32]. However, other variants of DI particles are also recognised including segments containing non-sense or other lethal mutations [33]. Accumulating experimental evidence demonstrates that not only is an IAV population comprised of a large proportion of virions that express an incomplete set of functional viral proteins, but that these "semi-infectious" virions are in the overwhelming majority, comprising up to 90% of the population depending on the strain. These defective viruses are capable of at least single-round infection [33] with the implication that an HA RNA segment with an HACS containing a nonsense mutation could still be packaged into a virion, infect a new host cell, and be replicated by virtue of complementation. It follows that frameshift mutations in the HACS may be restored by additional slippage in the first round of vRNA to cRNA transcription, or indeed subsequent vRNA to cRNA transcription or cycles thereof. Whether the RdRp slippage in the HACS occurs at the vRNA to cDNA stage or vice versa is unknown, similarly, it is not clear from the polyadenylation of mRNA whether

the stuttering occurs during cRNA synthesis or vRNA synthesis [28], but it was demonstrated that the RdRp is capable of reiterative copying of poly (U) as well as poly (A) tracts [28].

### *3.4 Length and composition of the connecting peptides in the quasispecies*

All field-derived LPAI H5 viruses have four amino acids in the connecting peptide (e.g. RETR) [20], but HPAI H5 and H7 field isolates with connecting peptides that vary from five up to eleven amino acids have been isolated, with seven or eight amino acids being the standard [34]. Does a longer connecting peptide confer a fitness advantage, and if the ostrich strain of this study had replicated further in ostriches would a six- seven- or eight amino acid connecting peptide insert eventually succeed as the master sequence? This remains unclear, but Horimoto and Kawaoka [21] mutated a H5N9 HPAI HACS to contain additional basic amino acids and that found that the longer mutant with eight amino acids in the connecting peptide had reduced cleavability compared to the parental sequence that contained only six basic amino acids. They concluded that different strains may have different thresholds for the length of insertions tolerated, and that when the population reaches equilibrium the beneficial effects become diluted. Regardless of length, the composition of the HACS motif is important and even slight sequence differences are capable of shifting the virus' dependence to an alternative protease, for example MSPL/ TMPRSS13 was experimentally demonstrated to preferentially cleave the connecting peptide sequence KKKR over furin, which has a preference for RKKR [6]. Effectively, the proteolytic cleavage specificity of HACS adds a further layer of selection pressure at the species, organ or cell type level depending on the protease expression profile. The variety of HACS motifs in the ostrich tissue represent those in pooled organs but it would have been interesting to determine whether different subsets were expressed in different tissues. Nonetheless it may be concluded that the master sequence here, PQRRKKRGLF, represents the substrate that is cleaved by a broad subset of proteases present in the tissues examined.

### *3.5 LPAI to HPAI: correlation of quasispecies data with observations of* **in vivo** *clinical studies*

In a seminal study, Ito and co-workers [7] passaged an LPAI H5 virus (PQRETRGLF) twenty four times in chick air sacs to adapt the virus for replication, followed by a further five passages through chick brains, ultimately producing viruses with an HACS motif of PQRRKKRGLF. For the first 18 passages the HACS retained the default LPAI motif, but by the end of the 24$^{th}$ passage in air sacs the HACS sequence had mutated to PQREKRGLF. This virus retained the avirulent phenotype indicated by a lack of clinical signs, restriction to growth in the trachea, and the requirement for exogenous trypsin in order to form plaques in cell culture. The REKR mutation in the connecting peptide was the result of a C to A mutation in the cDNA sequence AGA$^{R}$GAA$^{E}$AAA$^{K}$AGA$^{R}$ (amino acids in superscript; [7]). This same C to A mutation is found in the viral quasispecies of the present study (Table 1) and evidently the mutation is highly conserved across the quasispecies. Interestingly, in July 2014 an unrelated H5N2 strain that had been circulating in ostrich flocks for several weeks was identified by PCR and conventional Sanger sequencing to contain a PQREKRGLF motif (M. Romito, personal communication). Fortunately in that case the control measures could be applied before an HPAI strain emerged. The mutation to PQREKRGLF in the HACS of H5 viruses may thus be the first step in the conversion of LPAI to HPAI. It is not, however a sequence that is frequently detected in the field as only nine H5 isolates, all derived from poultry, contained this sequence out of 3140 in the public sequence database [34], and it may therefore represent a transient state which the results of Ito et al [7] appear to support.

Another correlation between the ostrich H5 virus quasispecies and Ito's passage experiment relates to the insertion of an additional amino acid in the connecting peptide, viz. PQRKKRGLF to PQRRKKRGLF [7]. With reference to the quasispecies distribution in Table 1, insertion of the additional arginine-encoding AGA codon in the hypothetical LPAI cDNA sequence CCT $^{P}$CAA$^{Q}$A**G**A $^{R}$**G**AA$^{E}$ ACA$^{T}$ AGA$^{R}$ GGT$^{G}$CTA$^{L}$TTT$^{F}$ cannot be achieved by the duplication of the preceding underlined

arginine codon, because evidence for a third guanine residue in the proximity is lacking. Instead, the arginine pair appears to be formed by the insertion of an additional adenine residue as underlined: $^{P}$CAA$^{Q}$A**G**A $^{R}$ A**G**A$^{R}$. This insertion is found across the quasispecies, as well as in the HPAI sequences described by Ito et al [7]. Point insertions are uncommon in IAV genomes [20], but they are readily visible in NGS sequence data when sequence reads are mapped against a reference (Supplemental Fig.). The selection of this adenine in the HACS may represent the first example of a point insertion in the IAV genome that is under positive selection pressure. This point insertion introduces a frameshift in the HA0 protein, but as discussed in section 3.3, this is not immediately detrimental or a limiting factor in the context of the entire population.

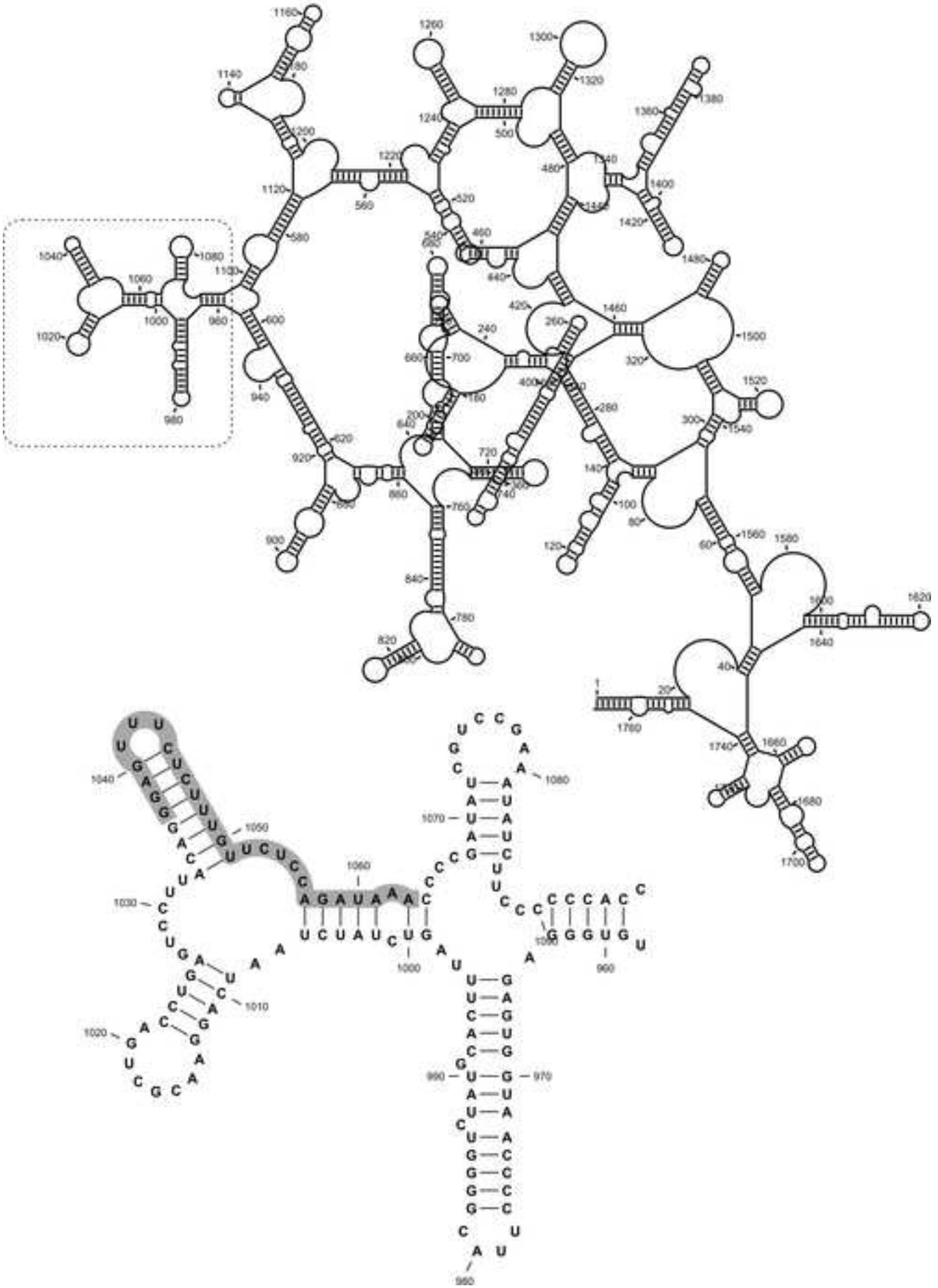
### *3.6 Comparative RNA structures and the conversion from LPAI to HPAI*

In view of the importance of RNA secondary structure in the synthesis of mRNA poly (A) tails [28-30], and evidence from the quasispecies that a similar process occurs in the HACS, the RNA structures of LPAI and HPAI H5 were compared. In the LPAI H5 structure (Fig 1a) the HACS spans base-paired RNA structures, but in the HPAI form (Fig 1b), the structure is altered so that the bulk of the connecting peptide-encoding region is shifted to a single-stranded bulge adjacent to a much smaller hairpin loop. The single-stranded bulge is the incorporation site for additional uracils, but it remains unclear which proximal secondary structure provides the steric hinderance to cause the RdRp to stutter. In the previous section it was postulated that the PQREKRGLF HACS motif may represent a transient form in the mutation of LPAI to HPAI. This structure is presented (Fig 1c), and similar to the classical LPAI sequence, the HACS is located in a base-paired RNA region, a hairpin loop in this case. Our current understanding of RNA secondary structure has not yet advanced to know whether or how base-paired structures affect RdRp fidelity during replication.

In section 3.5, evidence from the quasispecies, supported by published clinical studies pointed to the insertion of an additional adenine residue (uracil in the vRNA) as an early event in the formation of an arginine pair in the HACS of this particular virus. When the RNA sequence for PQREKRGLF HACS

Figure 1a





# (b) HPAI H5: PQRRKKRGLF

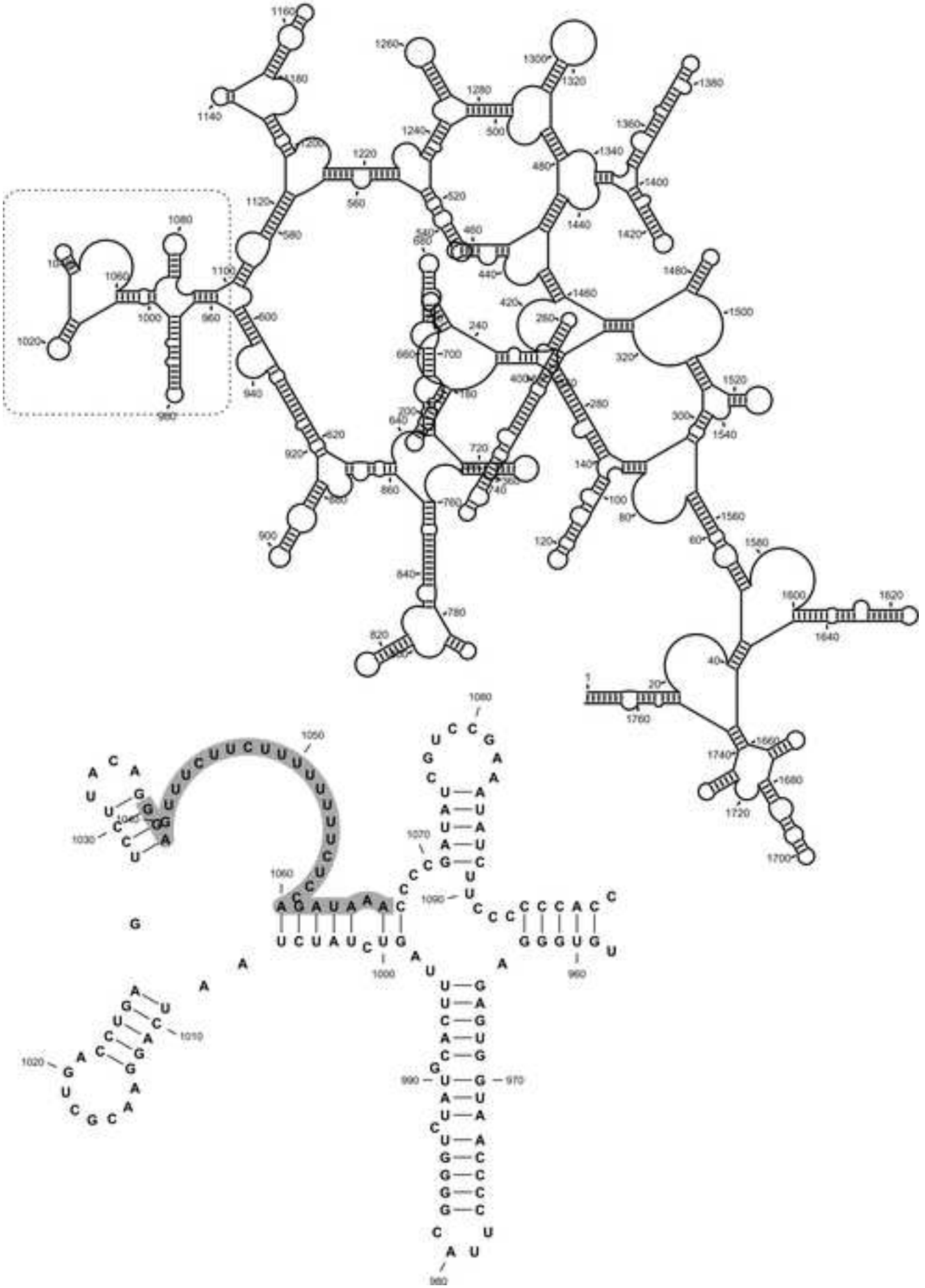
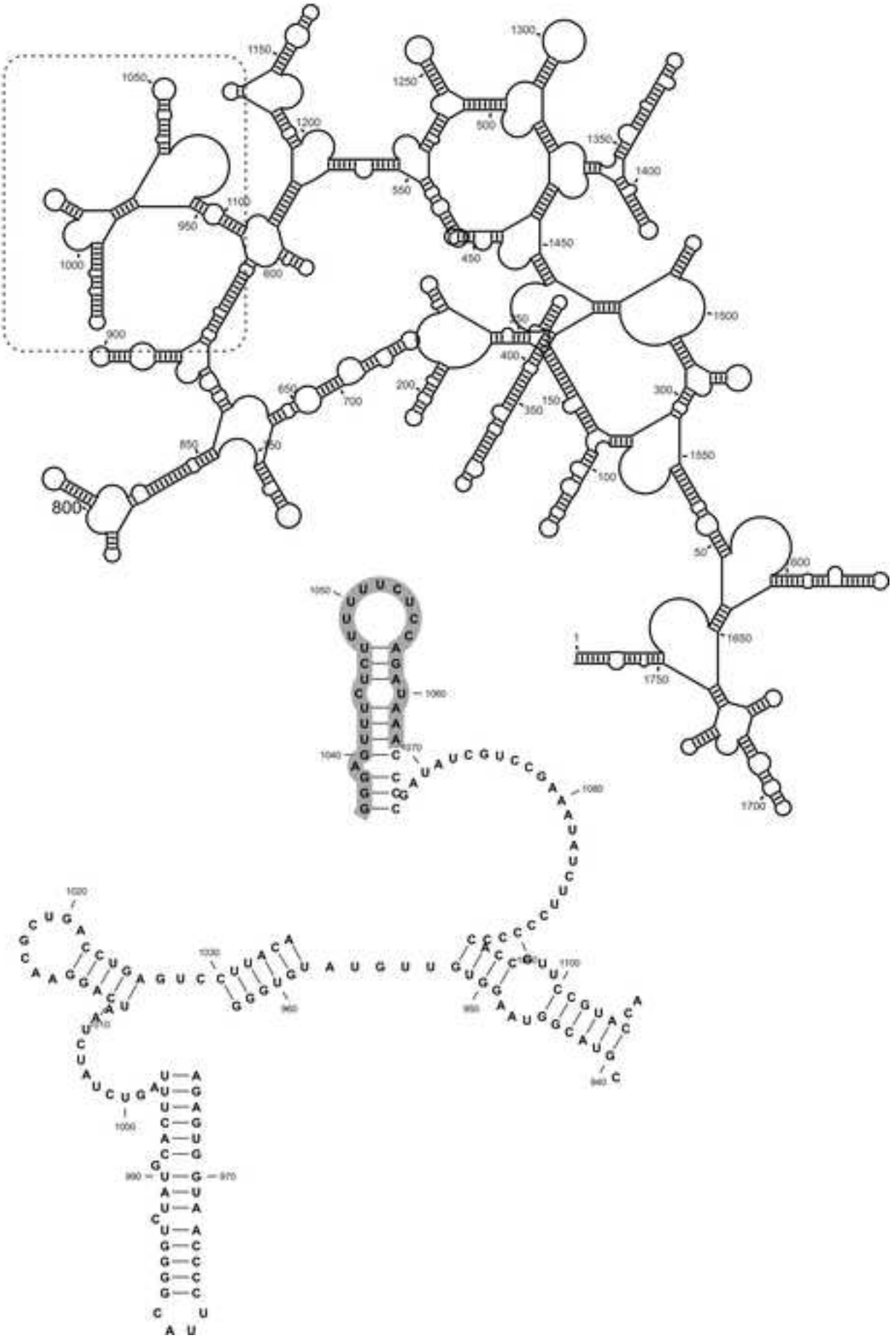
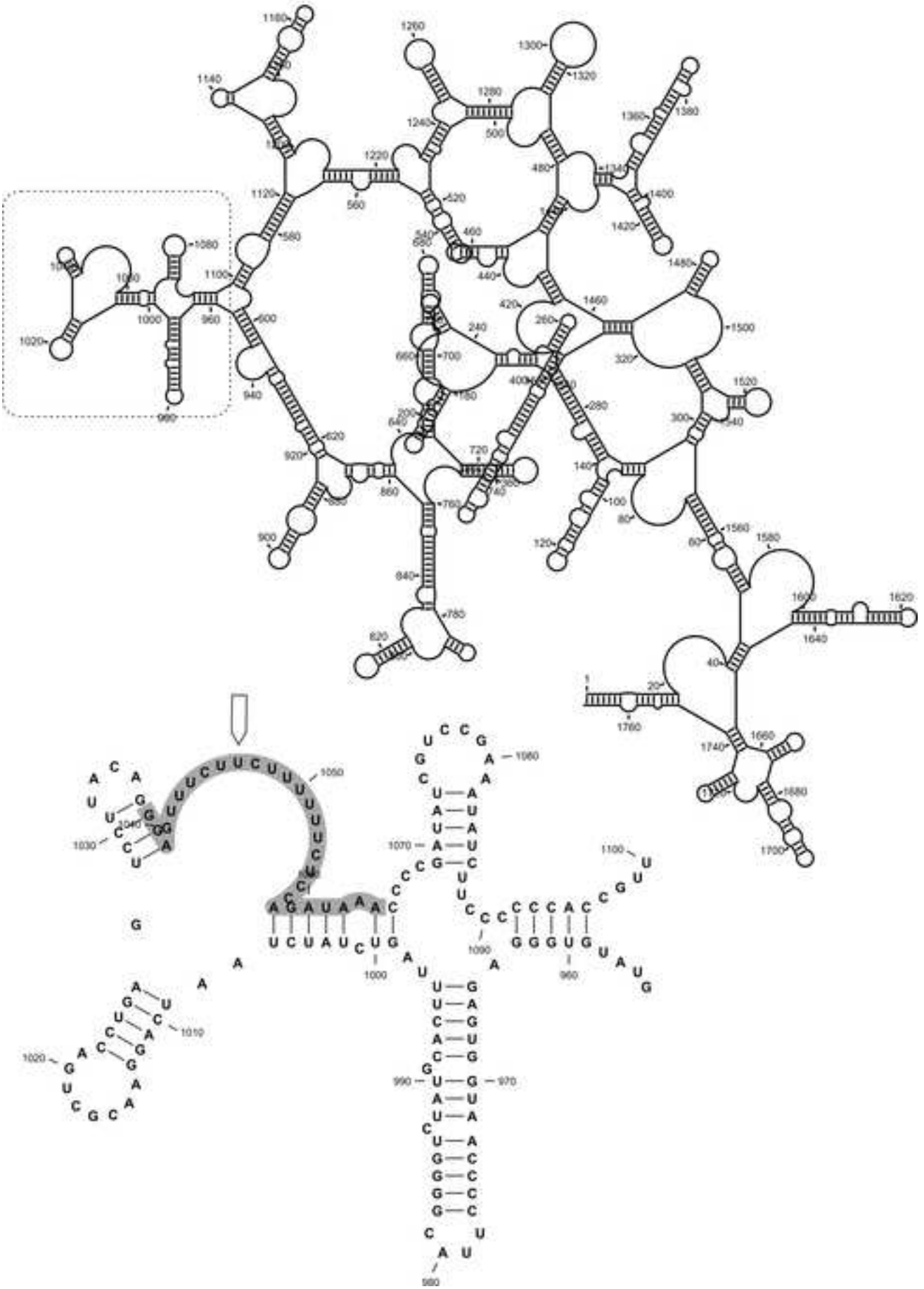


Figure 1c

# (c) LPAI H5: PQREKRGLF



(d) (PQRRKKRSI)



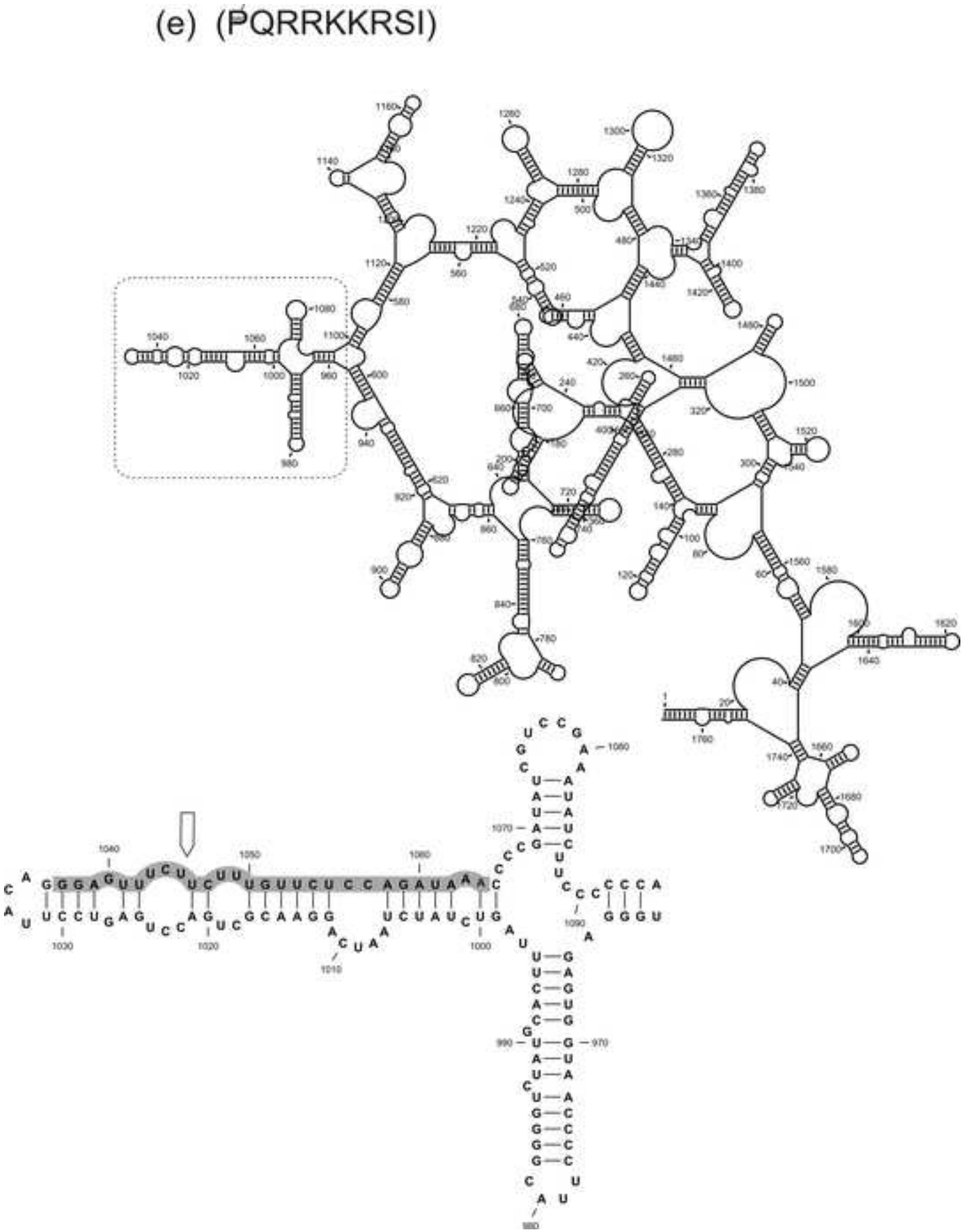



**Fig. 1.** Predicted RNA structures for segment 4 encoding the haemagglutinin protein of the H5 virus. The location of the HACS is highlighted and the arrow in 1(d) and (e) indicates the location of U insertion

was modified with this uracil insertion and refolded (Fig 1d), surprisingly, a structure almost identical to the HPAI sequence was obtained, with the HACS located on a single-stranded bulge. Insertion of this additional uracil into the RNA encoding the classical PQRETRGLF HACS did not have the same effect as the HACS remained located in base-paired RNA structure (Fig 1e). This seems to support the theory that the PQREKRGLF motif is an intermediary in the conversion of LPAI to HPAI. The formation of the HPAI-like secondary structure implies that the HACS is primed for RdRp slippage. Figure 2 is presented to summarise these steps and collates the findings of the previous section in defining the process of conversion from LPAI to HPAI.

Step one (2a) involves the random mutation of RETR in the connecting peptide to REKR, and this altered motif provides a biological advantage as it soon pervades the quasispecies. The viruses containing this sequence remain phenotypically avirulent [7], and despite a change in RNA conformation, the HACS remains located within base-paired RNA structure. The second step (2b) involves the selection of a single misincorporated uridine residue (adenine in the cDNA), and its insertion has two consequences. Firstly, in (2c), the conformation of the RNA switches to an HPAI-like structure that is primed for RdRp slippage, and secondly, the insertion results in a frameshift in HA0. This frameshift mutation is visible in the quasispecies and is represented by the peptide sequence PQRRKKRSI, present in the quasispecies at an above-average frequency of 0.88% (Supplemental Fig.). In a subsequent round of replication, the insertion of an additional uracil by RdRp slippage in the bulge (2d) produces another frame-shifted HA0 sequence encoding the HACS motif PQRRKKEVY. This sequence was present in the quasispecies at a frequency of 6.49% in the ostrich tissue. A second uracil misincorporation in the connecting peptide in the next round of replication restores the HA0 reading frame (2e), with an HACS that is consistent with the master sequence.

Computer-predicted RNA structures can differ dramatically among various IAV strains [35] and a wide variety of H5 HPAI HACS motifs have been recorded [34]. The mechanism described above may

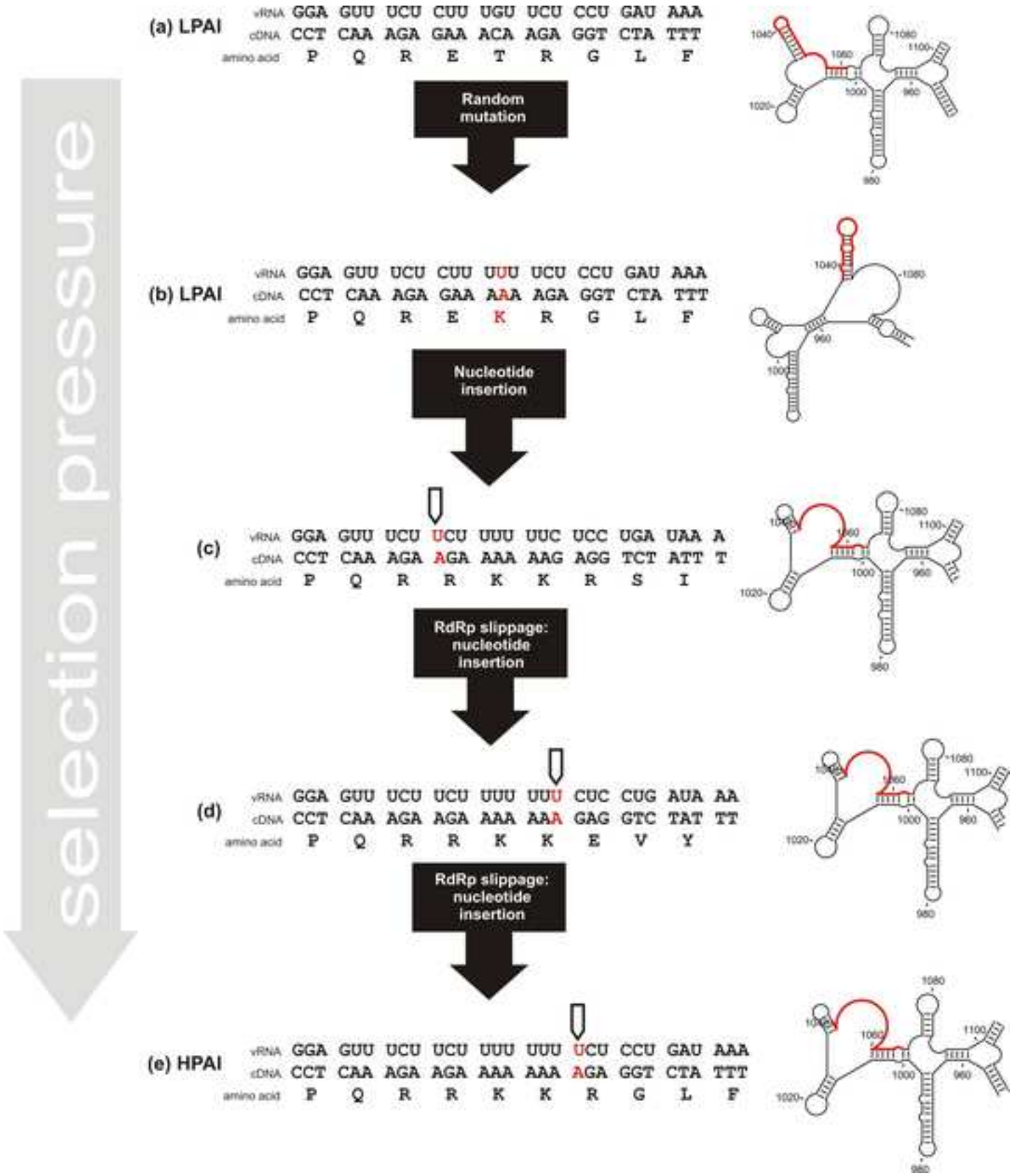



**Fig. 2.** The proposed mechanism for the conversion of LPAI to HPAI for strain A/ostrich/South Africa/2114/2011. In the LPAI progenitor (2a), a cytosine to uracil mutation under strong positive selection converts the HACS motif to PQREKRGLF (2b). The insertion of a uracil in (2c) switches the RNA conformation in the HACs to an HPAI-like structure. Subsequent RdRp slippage and uracil incorporation in the connecting peptide (2d, 2e) restores the HPAI motif

not be the exact mechanism that all H5 (or H7) strains follow in their conversion, but it identifies principles that may be common to all. This, however remains be investigated on a case-by-case basis. Therefore, the mechanism described here is not proposed as a general model, but rather a plausible explanation for how HPAI emerged in this specific case. In the course of conducting systematic nucleotide substitutions followed by RNA-refolding (data not shown), some other mutations, for example $U^{1045}\rightarrow A$, were determined to shift the RNA structure of the classic PQRETRGLF sequence directly to the HPAI RNA conformation, without the requirement for the RE<u>K</u>R mutation and uracil insertion, thereby presenting an alternative pathway. These mutations were not however evident in the quasispecies analysed here, and were therefore not considered the pathway in this particular case.

## 4. Discussion

The exact mechanism of conversion of LPAI to HPAI in terrestrial poultry has not been experimentally demonstrated but it is generally accepted that that RNA secondary structure and polymerase slippage are involved in the generation of multi-basic HACSs [11, 20, 22, 35]. Here, the quasispecies at the HACS of HPAI H5-infected ostrich tissues revealed by direct NGS was examined. The first direct evidence for the slippage of the RdRp in the connecting peptide region is provided, but this would benefit from further experimental demonstration. The dominance of critical mutations in the quasispecies led to the theory that the PQREKRGLF motif in H5 viruses is a transient LPAI precursor to HPAI, and is supported by the results of published clinical studies [7]. *In silico* RNA folding predicted that this transient intermediary plays an important role in subsequent events that relocate the HACS from base-paired RNA regions to a single-stranded bulge, thereby priming the connecting peptide region for RdRp slippage. The RNA secondary structure unique to H5 and H7 that provides steric hinderance to the RdRp during vRNA/ cRNA replication was not identified here. Although it doesn't provide an encompassing model for LPAI to HPAI conversion for H5 and H7 IAV,

this study identifies principles that may be common in the conversion of all strains. Most pertinently, this analysis reveals how heavily reliant IAV is on stochastic events to generate HPAI from an LPAI precursor, and reaffirms the conclusions of other studies that IAVs effectively exist less as a population of intact virus than a swarm of complementation-dependent, semi-infectious virions [33].

The stochasticity of the required steps in the mutation of LPAI to HPAI explains something of the timing of emergence of HPAI in the field: HPAI doesn't emerge immediately after LPAI has been introduced into susceptible poultry or indeed at any defined interval. For example, in the Pennsylvanian H5N2 outbreak in 1983 the virus took six months to become highly virulent in chickens [21]. In Central Mexico in 1995, HPAI appeared following an estimated fifteen months of the circulation of an LPAI precursor [11], and in the Italian 1999-2000 H7N1 epidemic the LPAI progenitor was detected in turkeys nine months prior to the emergence of the HPAI virus [13]. The H5N2 virus in this study is estimated to have circulated for at least four months in ostrich flocks before the first HPAI strain emerged [13], and more recently in the USA likely progenitors of H7N8 HPAI turkey outbreak viruses were detected in wild waterfowl only two months prior [15]. How long it takes for HPAI to emerge from LPAI in poultry is therefore a question that cannot be answered with any accuracy.

Yet another unanswered question is why is the ability to mutate the H5 and H7 HACS is restricted to specific terrestrial avian species (e.g. chickens, turkeys and ostriches) since the emergence of HPAI from a LPAI progenitor has never been demonstrated in aquatic birds, especially domesticated ducks that are intensively farmed. Host interactions are vital at every stage of the IAV life cycle and the virus depends on cellular factors to complete its replication cycle [36]. More than three hundred host proteins co-immuno precipitate with IAV [37] and several of these cellular factors associate with the individual components of the RdRp complex to enhance viral RNA replication by various mechanisms. These cellular factors include BAT1, Hsp90, IREF-1/MCM, Tat-SF1, the large subunit of cellular polymerase II, cellular transcription repressor *DR1,* RNA-binding protein NXP2/MORC3 ,

DnaJA1/Hsp40 and ANP32A [37-47]. The involvement of a species-specific *cis*-acting viral genomic replication factor in RdRp fidelity and slippage at the HACS is likely yet unexplored. Identifying the host-specific factor/s involved in the generation of HPAI is important, not only for a better understanding of IAV biology but also for screening other species for their biological potential to produce HPAI.

**Compliance with Ethical Standards**

This work was supported by the National Research Foundation [grant #82392]. The author declares that no conflict of interest exists.

**References**

1. Huang RTC, Wahn K, Klenk HD, Rott R (1980) Fusion between cell membranes and liposomes containing the glycoprotein of influenza virus. Virology 104: 294-302.
2. Webster RG, Rott R (1987) Influenza virus A pathogenicity: the pivotal role of hemagglutinin. Cell 50: 665-666.
3. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y (1992) Evolution and ecology of influenza A viruses. Microbiology Reviews 56(1): 152-179.
4. Garten W, Klenk HD (2008) Monographs in Virology In: Klenk HD, Matrosovich MN, Stech J (eds) Avian influenza, Vol. 27, Karger, pp 156–167.
5. Fouchier RA, Munster V, Wallensten A, Bestebroer TM, Herfst S, Smith D, Rimmelzwaan GF, Olsen B, Osterhaus AD (2005) Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. Journal of Virology 79(5): 2814-2822.
6. Kido H, Okumura Y, Takahashi E, Pan HY, Wang S, Yao D, Yao M, Chida J, Yano M (2012) Role of host cellular proteases in the pathogenesis of influenza and influenza-induced multiple organ failure. Biochimica et Biophysica Acta 1824(1): 186-194.

7. Ito T, Goto H, Yamamoto E, Tanaka H, Takeuchi M, Kuwayama M, Kawaoka Y, Otsuki K (2001) Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. Journal of Virology 75(9): 4439-4443.
8. Stech O, Veits J, Weber S, Deckers D, Schröer D, Vahlenkamp TW, Breithaupt A, Teifke J, Mettenleiter TC, Stech J (2009) Acquisition of a polybasic hemagglutinin cleavage site by a low-pathogenic avian influenza virus is not sufficient for immediate transformation into a highly pathogenic strain. Journal of Virology 83(11): 5864-5868.
9. Banks J, Speidel EC, McCauley JW, Alexander DJ (2000) Phylogenetic analysis of H7 haemagglutinin subtype influenza A viruses. Archives of Virology 145: 1047-1058.
10. Bashiruddin JB, Gould AR, Westbury HA (1992) Molecular pathotyping of two avian influenza viruses isolated during the Victoria 1976 outbreak. Australian Veterinary Journal 69(6): 140-142.
11. García M, Crawford JM, Latimer JW, Rivera-Cruz E, Perdue ML (1996) Heterogeneity in the haemagglutinin gene and emergence of the highly pathogenic phenotype among recent H5N2 avian influenza viruses from Mexico. Journal of General Virology 77: 1493-1504.
12. Berhane Y, Hisanaga T, Kehler H, Neufeld J, Manning L, Argue C, Handel K, Hooper-McGrevy K, Jonas M et al. (2009) Highly pathogenic avian influenza virus A (H7N3) in domestic poultry, Saskatchewan, Canada, 2007. Emerging Infectious Diseases 15(9): 1492-1495.
13. Abolnik C, Olivier AJ, Grewar J, Gers S, Romito M (2012) Molecular analysis of the 2011 HPAI H5N2 outbreak in ostriches, South Africa. Avian Diseases 56(4 Suppl): 865-879.
14. Monne I, Fusaro A, Nelson MI, Bonfanti L, Mulatti P, Hughes J, Murcia PR, Schivo A, Valastro V et al. (2014) Emergence of a highly pathogenic avian influenza virus from a low-pathogenic progenitor. Journal of Virology 88(8): 4375-4388.
15. Killian ML, Kim-Torchetti M, Hines N, Yingst S, DeLiberto T, Lee DH (2016) Outbreak of H7N8 low pathogenic avian influenza in commercial turkeys with spontaneous mutation to highly pathogenic avian influenza. *Genome Announcements* 4(3): pii: e00457-16

16. Li SQ, Orlich M, Rott, R (1990) Generation of seal influenza virus variants pathogenic for chickens, because of hemagglutinin cleavage site changes. Journal of Virology 64(7): 3297-3303.
17. Maurer-Stroh S, Lee RT, Gunalan V, Eisenhaber F (2013) The highly pathogenic H7N3 avian influenza strain from July 2012 in Mexico acquired an extended cleavage site through recombination with host 28S rRNA. Virology Journal 10: 139.
18. Pasick J, Handel K, Robinson J, Copps J, Ridd D, Hills K, Kehler H, Cottam-Birt C, Neufeld J et al. (2005) Intersegmental recombination between the haemagglutinin and matrix genes was responsible for the emergence of a highly pathogenic H7N3 avian influenza virus in British Columbia. Journal of General Virology 86(3): 727-731.
19. Suarez DL, Senne DA, Banks J, Brown IH, Essen SC, Lee CW, Manvell RJ, Mathieu-Benson C, Moreno V et al. (2004) Recombination resulting in virulence short in avian influenza outbreak, Chile. Emerging Infectious Diseases 10: 693-699.
20. Perdue ML, García M, Senne D, Fraire M (1997) Virulence-associated sequence duplication at the hemagglutinin cleavage site of avian influenza viruses. Virus Research 49(2): 173-186.
21. Horimoto T, Kawaoka Y (1997) A possible mechanism for selection of virulent avian influenza A viruses in 14-day-old embryonated eggs. Journal of Veterinary Medical Science 60(2): 273-275.
22. Spackman E, Senne DA, Davison S, Suarez DL (2003) Sequence analysis of recent H7 influenza viruses associated with three different outbreaks in commercial poultry in the United States. Journal of Virology 77: 13399-13402.
23. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95-98.
24. Desselberger U, Racaniello VR, Zazra JJ, Palese P (1980) The 3' and 5'-terminal sequences of influenza A, B and C virus RNA segments are highly conserved and show partial inverted complementarity. Gene 8(3): 315-328.

25. Zuker M, Stiegler, P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9(1):133-148.
26. Nobusawa E, Sato K (2006) Comparison of the mutation rates of human influenza A and B viruses. Journal of Virology 80(7): 3675-3678.
27. Luo GX, Luytjes W, Enami M, Palese P (1991) The polyadenylation signal of influenza A virus RNA involves a stretch of uredines followed by the RNA duplex of the panhandle structure. Journal of Virology 65: 2861-2867.
28. Poon LL, Pritlove DC, Fodor E, Brownlee GG (1999) Direct evidence that the poly(A) tail of influenza A virus mRNA is synthesized by reiterative copying of a U track in the virion RNA template. Journal of Virology 73: 3473–3476.
29. Pritlove DC, Poon LLM, Devenish L, Leahy MB, Brownlee G (1999) A hairpin loop at the 5' end of influenza A virus virion RNA is required for synthesis of Poly(A) mRNA *in vitro*. Journal of Virology 73(3): 2109-2114.
30. Zheng H, Lee HA, Palese P, Garcia-Sastre A (1999) Influenza A virus RNA polymerase has the ability to stutter at the polyadenylation site of a viral RNA template during RNA replication. Journal of Virology 73(6): 5240-5243.
31. García-Sastre A, Palese P (1993) Genetic manipulation of negative-strand RNA virus genomes. Annual Reviews in Microbiology 47: 765-790.
32. Laske T, Heldt FS, Hoffman H, Frensing T, Reichl U (2016) Modelling the intracellular replication of influenza A virus in the presence of defective interfering RNAs. Virus Research 213: 90-99.
33. Brooke CB, Ince WL, Wrammert J, Ahmed R, Wilson PC, Bennink JR, Yewdell JW (2013) Most influenza A virions fail to express at least one essential viral protein. Journal of Virology 87(6): 3155-3162.

34. Luczo JM, Stambas J, Durr PA, Michalski WP, Bingham J (2015) Molecular pathogenesis of H5 highly pathogenic avian influenza: the role of the haemagglutinin cleavage site motif. Reviews in Medical Virology 25: 406-430.
35. Perdue ML, Suarez DL (2000) Structural features of the avian influenza virus hemagglutinin that influence virulence. Veterinary Microbiology 274: 77-86.
36. Hsu SF, Su WC, Jeng KS, Lai MM (2015) A host susceptibility gene, DR1, facilitates influenza A virus replication by suppressing host innate immunity and enhancing viral RNA replication. Journal of Virology 89(7): 3671-3682.
37. Watanabe T, Kawakami E, Shoemaker JE, Lopes TJ, Matsuoka Y, Tomita Y, Kozuka-Hata H, Gorai T, Kuwahara T et al. (2014) Influenza virus-host interactome screen as a platform for antiviral drug development. Cell Host Microbe 16(6): 795-805
38. Momose F, Naito T, Yano K, Sugimoto S, Morikawa Y, Nagata K (2002) Identification of Hsp90 as a stimulatory host factor involved in influenza virus RNA synthesis. Journal of Biological Chemistry 277(47): 45306-45314.
39. Mayer D, Molawi K, Martínez-Sobrido L, Ghanem A, Thomas S, Baginsky S, Grossmann J, García-Sastre A, Schwemmle M (2007) Identification of cellular interaction partners of the influenza virus ribonucleoprotein complex and polymerase complex using proteomic-based approaches. Journal of Proteome Research 6(2): 672-682.
40. Hao L, Sakurai A, Watanabe T, Sorensen E, Nidom CA, Newton MA, Ahlquist P, Kawaoka Y (2008) Drosophila RNAi screen identifies host genes important for influenza virus replication. Nature 454: 890–893.
41. Brass AL, Huang IC, Benita Y, John SP, Krishnan MN, Feeley EM, Ryan BJ, Weyer JL, van der Weyden L et al. (2009) The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. Cell 139: 1243–1254.

42. Shapira SD, Gat-Viks I, Shum BO, Dricot A, de Grace MM, Wu L, Gupta PB, Hao T, Silver SJ, Root DE, Hill DE, Regev A, Hacohen N (2009) A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. Cell 139: 1255–1267.

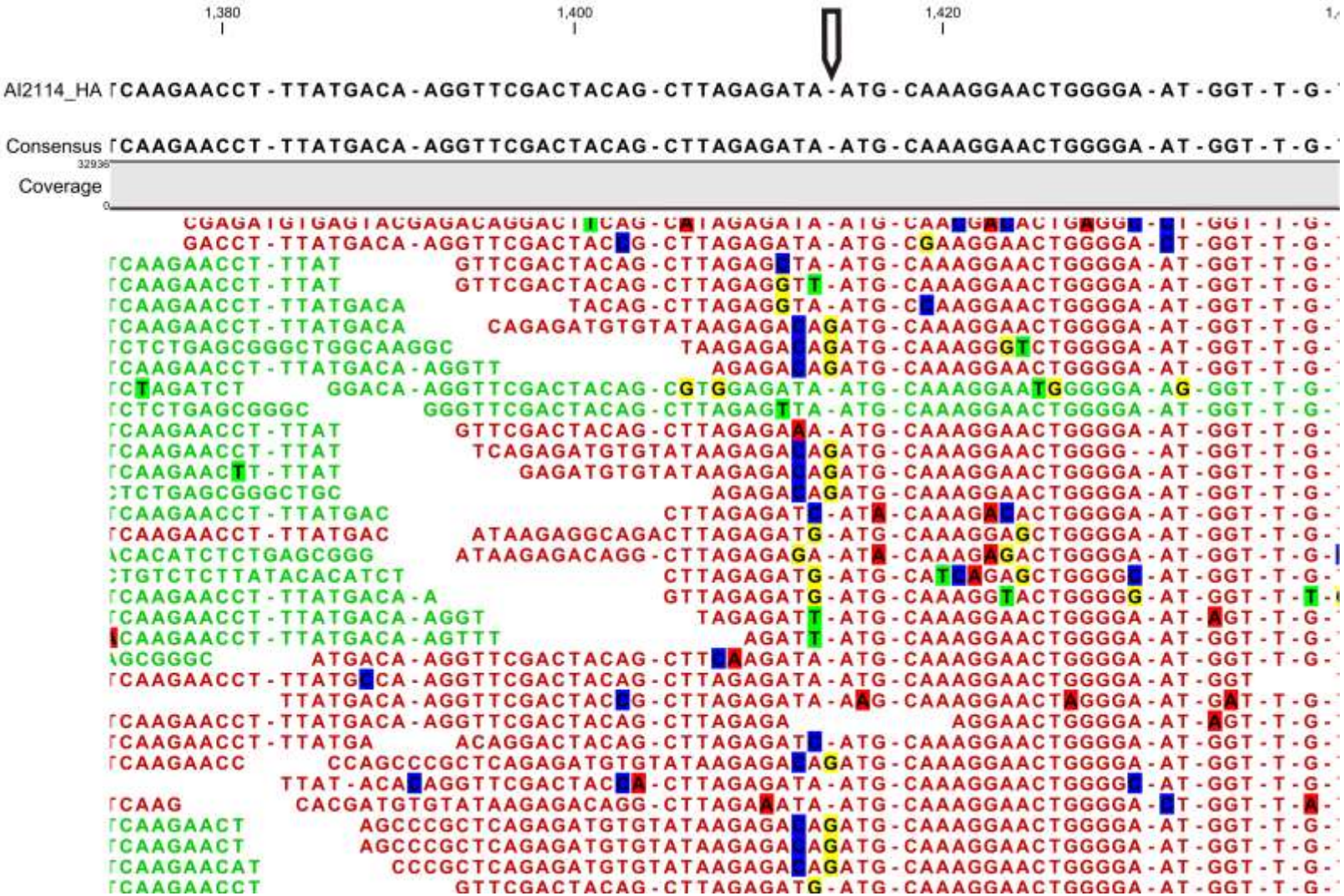
43. Karlas A, Machuy N, Shin Y, Pleissner KP, Artarini A, Heuer D, Becker D, Khalil H, Ogilvie LA et al.. (2010) Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. Nature 463: 818–822.

44. Konig R, Stertz S, Zhou Y, Inoue A, Hoffmann HH, Bhattacharyya S, Alamares JG, Tscherne DM, Ortigoza MB et al. (2010) Human host factors required for influenza virus replication. Nature 463: 813–817.

45. Cao M, Wei C, Zhao L, Wang J, Jia Q, Wang X, Jin Q, Deng T (2014) DnaJA1/Hsp40 is co-opted by influenza A virus to enhance its viral RNA polymerase activity. Journal of Virology 88(24): 14078-14089

46. Ver LS, Marcos-Villar L, Landeras-Bueno S, Nieto A, Ortín J (2015) The Cellular Factor NXP2/MORC3 Is a Positive Regulator of Influenza Virus Multiplication. Journal of Virology 89(19): 10023-10030.

47. Long JS, Giotis ES, Moncorgé O, Frise R, Mistry B, James J, Morisson M, Iqbal M, Vignal A et al. (2016) Species difference in ANP32A underlies influenza A virus polymerase host restriction. Nature 529: 101-104.




**Supplemental figure.** Insertions in the viral quasispecies are evident in NGS data when sequence reads are mapped to a reference. The arrow indicates the location of a guanine residue that is not present in the consensus sequence. (JPEG 4721 kb)