https://doi.org/10.1038/s44298-024-00029-1

Nucleotide sequence as key determinant driving insertions at influenza A virus hemagglutinin cleavage sites

Monique I. Spronken[1], Mathis Funk[1], Alexander P. Gultyaev[1], Anja C. M. de Bruin[1,2], Ron A. M. Fouchier[1] & Mathilde Richard[1] ✉

Highly pathogenic avian influenza viruses (HPAIVs) emerge from H5 and H7 low pathogenic avian influenza viruses (LPAIVs), most frequently upon insertions of nucleotides coding for basic amino acids at the cleavage site (CS) of the hemagglutinin (HA). The exact molecular mechanism(s) underlying this genetic change and reasons underlying the restriction to H5 and H7 viruses remain unknown. Here, we developed a novel experimental system based on frame repair through insertions or deletions (indels) of HAs with single nucleotide deletions. Indels were readily detected in a consensus H5 LPAIV CS at low frequency, which was increased upon the introduction of only one substitution leading to a longer stretch of adenines at the CS. In contrast, we only detected indels in H6 when multiple nucleotide substitutions were introduced. These data show that nucleotide sequence is a key determinant of insertions in the HA CS, and reveal novel insights about the subtype-specificity of HPAIV emergence.

Wild aquatic birds are the original reservoir of influenza A viruses, which are categorised by the antigenic properties of their hemagglutinin (HA) and neuraminidase (NA) surface glycoproteins. To date, 17 HA and nine NA subtypes have been detected in wild aquatic birds[1,2]. Influenza A viruses generally cause asymptomatic or mild infections in birds, and are referred to as low pathogenic avian influenza viruses (LPAIVs)[3]. Upon the introduction of LPAIVs of the H5 and H7 subtypes in terrestrial poultry, highly pathogenic avian influenza viruses (HPAIVs) may emerge. HPAIVs cause severe disease in poultry with mortality rates as high as 100%[4]. Besides these disastrous consequences on animal welfare and the poultry industry, spillover events to humans pose a continuous pandemic threat[5].

Post-translational cleavage of HA by host cell proteases is necessary for HA to be fusogenic and the virus to be infectious[1]. The HA proteins of LPAIVs have a monobasic cleavage site (CS) that is cleaved by trypsin-like proteases, predominantly expressed in the respiratory and intestinal tracts of birds[6,7]. In the vast majority of cases, the conversion from LPAIV to HPAIV is the result of the insertion of nucleotides coding for basic amino acids at the HA0 precursor protein CS, leading to a multi-basic cleavage site (MBCS)[8,9]. MBCSs are cleaved by furin-like proteases that are ubiquitously expressed, supporting systemic virus dissemination and severe disease in gallinaceous species[10]. Decades ago, different mechanisms of MBCS acquisition were identified upon the analysis of MBCS sequences of HPAIVs: (i) nucleotide substitutions and (ii) nucleotide insertions possibly due to either stuttering and/or backtracking of the influenza virus RNA-dependent RNA polymerase (RdRp) or, (iii) in the case of some H7 viruses, to non-homologous recombination with exogenous viral or host RNA[11–13]. The exact molecular mechanisms underlying MBCS acquisition via insertions remain unknown to date. Additionally, it remains unknown why HPAIVs have so far only evolved from H5 and H7 LPAIV precursors. Artificial introduction of an MBCS into non-H5/H7 HAs generally led to trypsin-independency in vitro[14–17], indicating that the absence of an MBCS in naturally occurring non-H5/H7 viruses is not due to incompatibility at the protein level. The restriction of HPAIVs to H5 and H7 subtypes might therefore be due to differences and/or constraints at the RNA level. RNA structures have been suggested to play a role in insertion generation at the HA CS[11,18–20]. Subtype-specific conserved secondary RNA stem-loop structures have been predicted at the HA CS[19,21]. However, RNA structure predictions of HAs from subtypes other than H5 and H7 have identified similar putative structures, suggesting that RNA structures might not be the only important factor that contributes to insertions in HA CSs[21]. The H5 and H7 CS sequences have been shown to stand out by their high purine content, suggesting that specific sequences and codon composition[11,18,20,22–25] may be important for MBCS acquisition and subtype restriction of HPAIV genesis[25].

[1]Department of Viroscience, Erasmus Medical Centre, Rotterdam, The Netherlands. [2]Institute for Biochemistry & Research Center for Emerging Infections and Zoonoses (RIZ), University of Veterinary Medicine Hannover, Hannover, Germany. ✉e-mail: m.richard@erasmusmc.nl

Research aimed to study the drivers of MBCS acquisition has been hampered by the fact that it is difficult to mimic this process in a laboratory setting, both in vitro and in vivo. H5 and H7 LPAIVs converted in rare occasions to HPAIVs and only when selection pressure for trypsin independency was applied[26–31], extensive passaging was performed[26,29,30,32,33], or non-basic amino acids at the CS were substituted to basic ones[22–24,34–36]. Some of these experiments were performed with virus isolates instead of clonal viruses, making the distinction between de novo insertions and the presence of viral quasispecies with mutated CSs impossible[26–33]. Other shortcomings were the low throughput and sensitivity as viruses with mutated CSs have to outcompete the wild-type (WT) virus in order to be reliably detected. To cope with these limitations, we developed a sensitive system with which indels at the HA CS can be readily detected. Single nucleotide deletions (SNDs) were introduced at the HA CS, leading to a non-functional HA protein. Upon use of these templates in reverse genetics, viruses with HAs that contain indels restoring the reading frame are under strong selective advantage without competition with a WT virus. Using this experimental system, we investigated the impact of nucleotide sequence at the HA CS on insertions and deletions (indels), in order to shed light on subtype-restriction of LPAIV to HPAIV conversion. We showed that nucleotide sequence was important in the acquisition of indels in the H5 and H6 CS region. Indels in the H5 CS were easily facilitated by a one nucleotide substitution increasing the length of the adenine/uracil (A/U)-stretch while indels in the H6 CS were only detected upon five or six nucleotide substitutions.

Results

Indels were readily detected at the H5 CS

We aimed to develop a novel experimental system with which insertions at the HA CS can be detected with a high sensitivity. SNDs were introduced, leading to a (−1) frameshift in the HA reading frame and thus changing the coding sequence. Upon use of these templates in reverse genetics, functional virus can only be produced if the frame of the HA protein is repaired by indels. This system ensures that rare indels that would not necessarily confer a selective advantage over the WT virus can be detected.

In order to study indels in an H5 with a LPAIV CS, the CS of the H5 A/Indonesia/5/2005 (A/Indo/5/05) HPAIV HA was mutated to match the H5 LPAIV amino acid and nucleotide consensus P4 to P1 RETR sequence ($H5_{RETR}$)[25] (Supplementary Fig. 1, Fig. 1a). SNDs were introduced into the $H5_{RETR}$ CS and the resulting plasmids were co-transfected with the remaining seven reverse genetics plasmids of A/Indo/5/05 as virus backbone. Each HA with a SND (HA_{SND}) was tested in three independent reverse genetics experiments. When virus was detected by observation of cytopathic effects and/or hemagglutination assay, the HA CS region was sequenced by Sanger sequencing to determine the nature of the indels which resulted in frame repair and recombinant virus production. Predicted RNA stem-loops (Supplementary Fig. 1) and positive-sense orientation are used throughout the manuscript to represent the data and describe the location and nature of the indels. Ten $H5_{RETR}$ HA_{SND}s were tested and virus was detected with only three HA_{SND}s in 13% (4/30) of the experiments (Fig. 1b). Two single A insertions in the A-stretch located at the 3' end of the loop, a four-nucleotide insertion at the 3' end and a two-nucleotide deletion in the A-stretch at the loop 5' end were detected (Fig. 1b, Supplementary Table 1a). These insertions led to tribasic and tetrabasic CSs (RKTR, REKR and REKKR; Supplementary Table 1a).

Next, SNDs were introduced in the HA of A/Indo/5/05 ($H5_{MBCS}$). Nineteen $H5_{MBCS}$ HA_{SND}s were tested. Virus was detected in either 2/3 or 3/3 experiments with 16 out of 19 HA_{SND}s, resulting in a total of 46 detected indels. Two-nucleotide deletions (16/46), single-nucleotide insertions (15/46) and insertions of more than one nucleotide (15/46) were observed (Fig. 1c). Indels were mostly observed in the loop of the predicted $H5_{MBCS}$ RNA structure, and only 2/46 of the indels were located at the top of the 5' side of the stem. Indels observed in the loop were mainly located in two regions: the large A-stretch at the 3' end of the loop (position 1057-1062) and the small A-stretch at the 5' end of the loop (position 1041-1043). Indels that were observed repeatedly and in multiple HA_{SND}s across the study were defined as indel pattern 1–4. Indel pattern 1 corresponded to a seven-nucleotide insertion (AAGAAAA) and represented 28% (13/46) of the observed indels in $H5_{MBCS}$ HA_{SND}s. Indel pattern 2 corresponded to a one A insertion in the A-stretch at the 3' end of the loop and was observed in 26% (12/46) of indels in $H5_{MBCS}$ HA_{SND}s. Indel pattern 3 corresponded to a two-nucleotide deletion in the A-stretch at the 5' end of the loop and was observed in 7% (3/46) of indels in $H5_{MBCS}$ HA_{SND}s. Finally, indel pattern 4 corresponded to the insertion of a single A in the A-stretch at the 5' end of the loop and was observed in 4% (2/46) of indels in $H5_{MBCS}$ HA_{SND}s. All other detected indels are shown in Supplementary Table 1a. As many indels were observed in homopolymers, SNDs were introduced in A, U, cytosine (C) and guanine (G) homopolymers in other regions of the influenza virus genome. Virtually no indels were detected (Supplementary Table 2), suggesting a peculiarity of the CS region.

Next, SNDs were introduced in the stem of the predicted RNA structures of $H5_{RETR}$ and $H5_{MBCS}$. No virus was detected in any of the 33 independent experiments (Supplementary Fig. 2a, b). This indicated that either indels did not occur or remained undetected because they did not lead to frame repair or led to non-functional HA proteins.

Taken together, these results show that indels were observed at the CS of both $H5_{RETR}$ and $H5_{MBCS}$ albeit at a much higher level in the $H5_{MBCS}$ HA. A striking difference between $H5_{RETR}$ and $H5_{MBCS}$ is the length of the A-stretch at the 3' end of the loop, which was three and six-nucleotides long, respectively. Moreover, indels were detected in $H5_{RETR}$ when the length of the 3' end loop A-stretch was increased from three to four or five (HA_{SND} Δ1046 and Δ1050; Fig. 1b), suggesting that the sequence of the 3' end of the loop might be important for indels at the HA CS.

Increasing the length of the A-stretch at the 3'end of the loop facilitated indels in H5

Our previous observations suggested that the presence of longer A-stretches might be an important driver of indels at the HA CS (Fig. 1). To further investigate the impact of sequence on indel generation, the length of the 3' end loop A-stretch was increased. Nucleotide substitutions were introduced in $H5_{RETR}$ at positions 1046, 1050 or both, which increased the A-stretch length to five, six or eight nucleotides, respectively, creating tri- and tetrabasic intermediate CS sequences (G1046A ($H5_{RKTR}$), C1050A ($H5_{REKR}$) and G1046A/C1050A ($H5_{RKKR}$); Fig. 2a).

Virus was detected in 67% (16/24), 88% (21/24) and 94% (17/18) of experiments using $H5_{RKTR}$, $H5_{REKR}$ and $H5_{RKKR}$ HA_{SND}s, respectively (Fig. 2b–d). This contrasted with data obtained with $H5_{RETR}$ (13%; 4/30), indicating that increasing the length of the 3' end loop A-stretch indeed promoted indels. Most of the observed indels corresponded to the previously identified indel patterns. Indel pattern 2 was most frequently observed, being identified in 73% (10/16), 67% (14/21) and 24% (4/17) of the viruses produced using $H5_{RKTR}$ and $H5_{REKR}$ and $H5_{RKKR}$ HA_{SND}s, respectively. A wider variation of indels was observed when $H5_{RKKR}$ HA_{SND}s were tested (Fig. 2d and Supplementary Table 1a). Several insertions of four, seven (indel pattern 1) and even 13 nucleotides were observed, all consisting of As and one to three Gs. All insertions of 4, 7 and more nucleotides appeared to be the result of duplications of neighbouring sequences. Interestingly, a deletion of two As (indel pattern 3) was consistently observed in $H5_{RKKR}$ HA_{SND} Δ1044, which only contained As in the loop, thereby suggesting that there may be a limit to the number of contiguous As that are tolerated at the HA CS. The HA CS sequence of some of the viruses could not be reliably determined by Sanger sequencing (indicated by a green wedge in Fig. 2), indicative of the presence of mixed virus populations. This was confirmed by performing TOPO-cloning of H5 CS PCR fragments and subsequent Sanger sequencing of six clones per fragment. A wide variety of

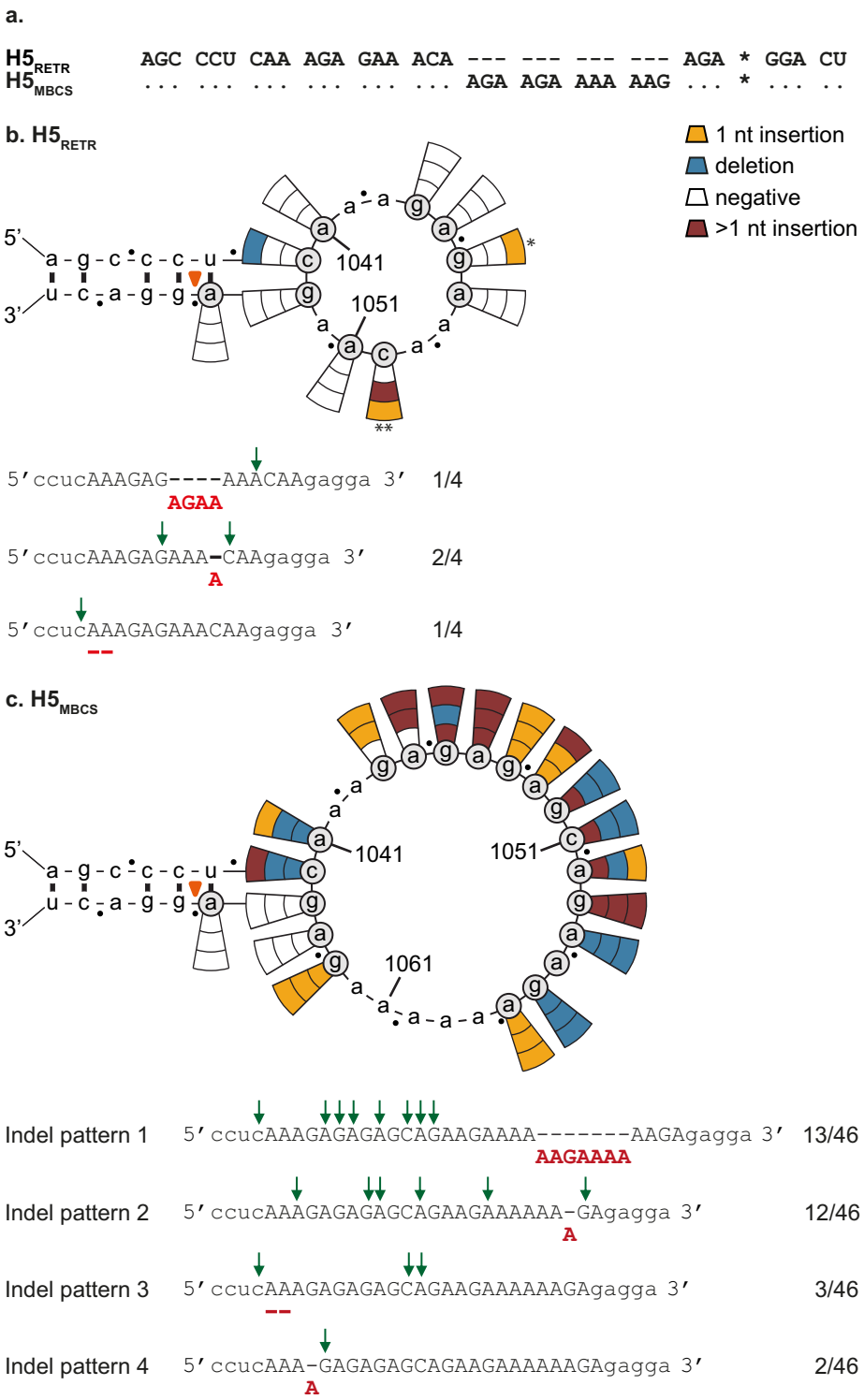


Fig. 1 | Higher indel detection was observed in the $H5_{MBCS}$ CS than in that of $H5_{RETR}$. The predicted RNA stem-loops serve as a basis to schematically represent the data. The black dots delineate codons. The orange arrow head indicates the start of the codon of the HA2 N-terminal glycine. Grey circles indicate the SNDs that were generated. Connecting pie slices show the results of three independent reverse genetics experiments, with the colour of the wedge indicating the nature of the observed indel, according to the legend at the top. The asterisks indicate results that are shown in multiple figures, with the number of asterisks identifying a given result. Observed indel patterns and corresponding frequencies are indicated below the structures, the other detected indels are shown in Supplementary Table 1a. The sequence of the loop and stem of the predicted RNA structures are shown with capital and lowercase letters, respectively. The green arrows indicate the location of each SND present in the HAs in which the corresponding indel pattern was detected. Sequences in all figures are indicated in cRNA positive sense orientation. **a** Alignment of $H5_{RETR}$ and $H5_{MBCS}$ CS sequences shown in Fig. 1. The asterisk indicates the border between HA1 and HA2. Results from testing (**b**) $H5_{RETR}$ HA_{SND}s and (**c**) $H5_{MBCS}$ HA_{SND}s.

sequences was detected (Supplementary Table 1b). In 58% (35/60) of the clones, a frameshift in the HA protein was detected due to either the presence of the introduced SND (69%; 24/35) or an indel that did not lead to reading frame repair. The most frequently observed indel was a one A insertion which resulted in a R323K amino acid change. Additionally, larger insertions of 3, 5, 7 and even 33 As were observed. Together, these data suggested that tri- and tetrabasic intermediate CS motifs were more error-prone than the LPAIV $H5_{RETR}$ CS.

The reciprocal approach was then taken by decreasing the length of the A-stretch of the H5s for which the highest indel frequency was observed ($H5_{MBCS}$, $H5_{REKR}$ and $H5_{RKKR}$), through the introduction of G or C nucleotides leading to either silent or non-silent substitutions. In general, reducing the length of the A-stretch decreased indel frequency (Supplementary Figs. 3–6: see notes in Supplementary information). Furthermore, a higher deletion to insertion rate was detected in $H5_{MBCS}$ with decreased A-stretch length, mostly occurring in the middle and at the 5' end of the loop.

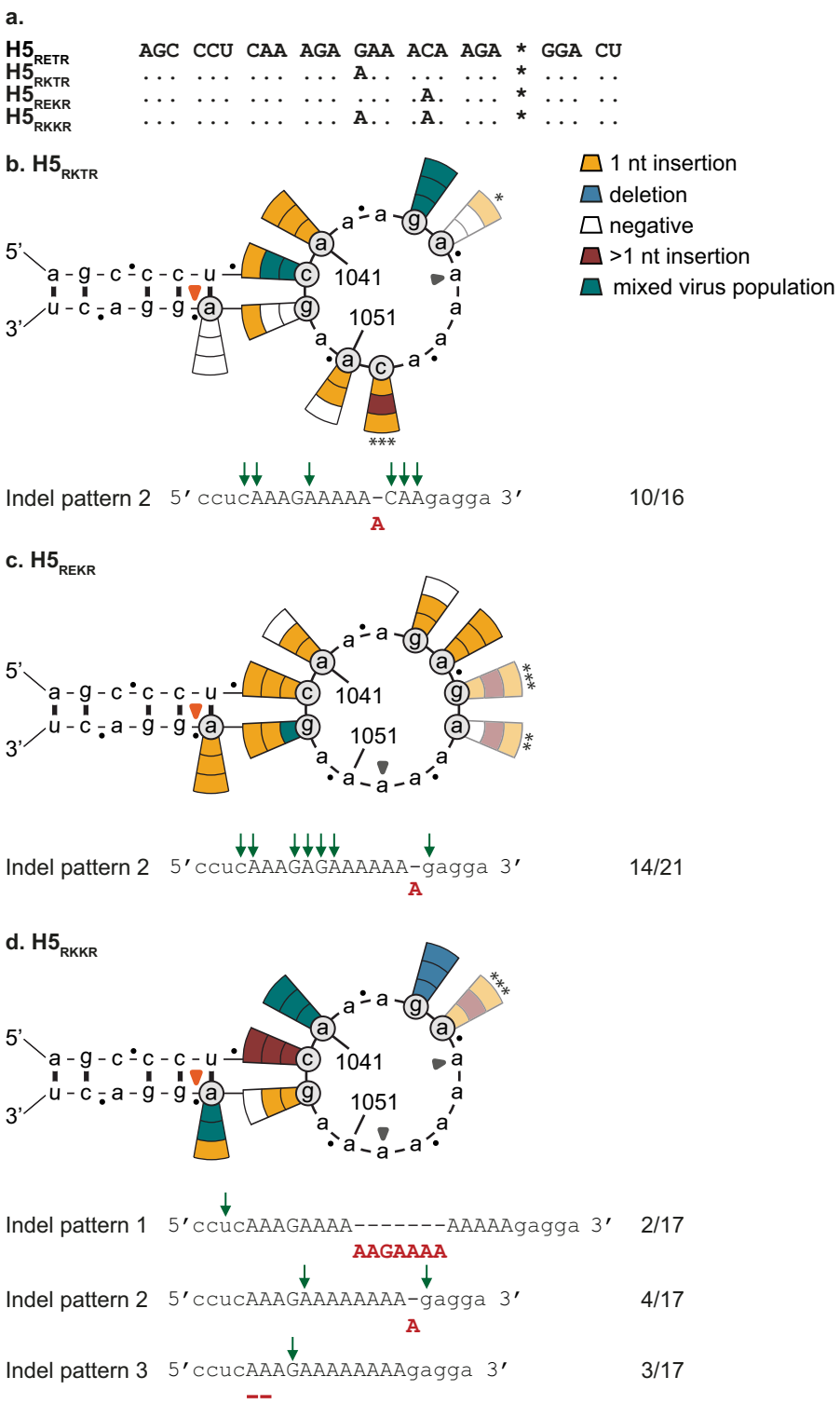


Fig. 2 | Increasing A homopolymer length facilitates indels at the H5 CS. Results are shown as described in the legend of Fig. 1. Grey closed arrow heads refer to the introduced nucleotide substitutions as compared to $H5_{RETR}$. A green wedge indicates that the HA sequence of the rescued virus could not be reliably determined by Sanger sequencing. Faded pie slices indicate that an HA with the same sequence was previously tested and show the data from the previous experiment, identified by the same number of asterisks. Observed indel patterns and corresponding frequencies are indicated below the structures, the other detected indels are shown in Supplementary Table 1a. **a** Alignment of H5 intermediate CS sequences shown in Fig. 2. The asterisk indicates the border between HA1 and HA2. Results from testing (**b**) $H5_{RKTR}$ HA_{SND}s, (**c**) $H5_{REKR}$ HA_{SND}s and (**d**) $H5_{RKKR}$ HA_{SND}s.

Taken together, these results revealed a positive association between the A-stretch length at the 3' end of the loop and indel occurrence at the H5 CS. This suggests that the sequence of the HA CS is a key determinant of indel generation and the nature of such indels.

Indels were observed at the H6 CS only after extensively changing the CS sequence

To gain further knowledge on subtype-restriction of MBCS acquisition, LPAIV $H6_{IETR}$ HA_{SND}s were tested in our experimental system. The HA of A/mallard/Sweden/81/2002 (H6N1; A/ml/Sw/81/02) was used as (i) no HPAIVs have so far evolved from H6 LPAIVs in nature, despite extensive circulation in poultry[37,38] and (ii) H6 and H5 belong to the same phylogenetic group1 (Fig. 3a). Moreover, the H6 CS region of A/ml/Sw/81/02 corresponds to the consensus CS sequence of all H6 LPAIVs, both at the amino acid and nucleotide level[25]. SNDs were introduced into $H6_{IETR}$ and reverse genetics experiments were performed using the H5 A/Indo/5/05 virus backbone. No virus was detected in any of the 42 experiments using $H6_{IETR}$ HA_{SND}s (Fig. 3b), reflecting the natural situation whereby no

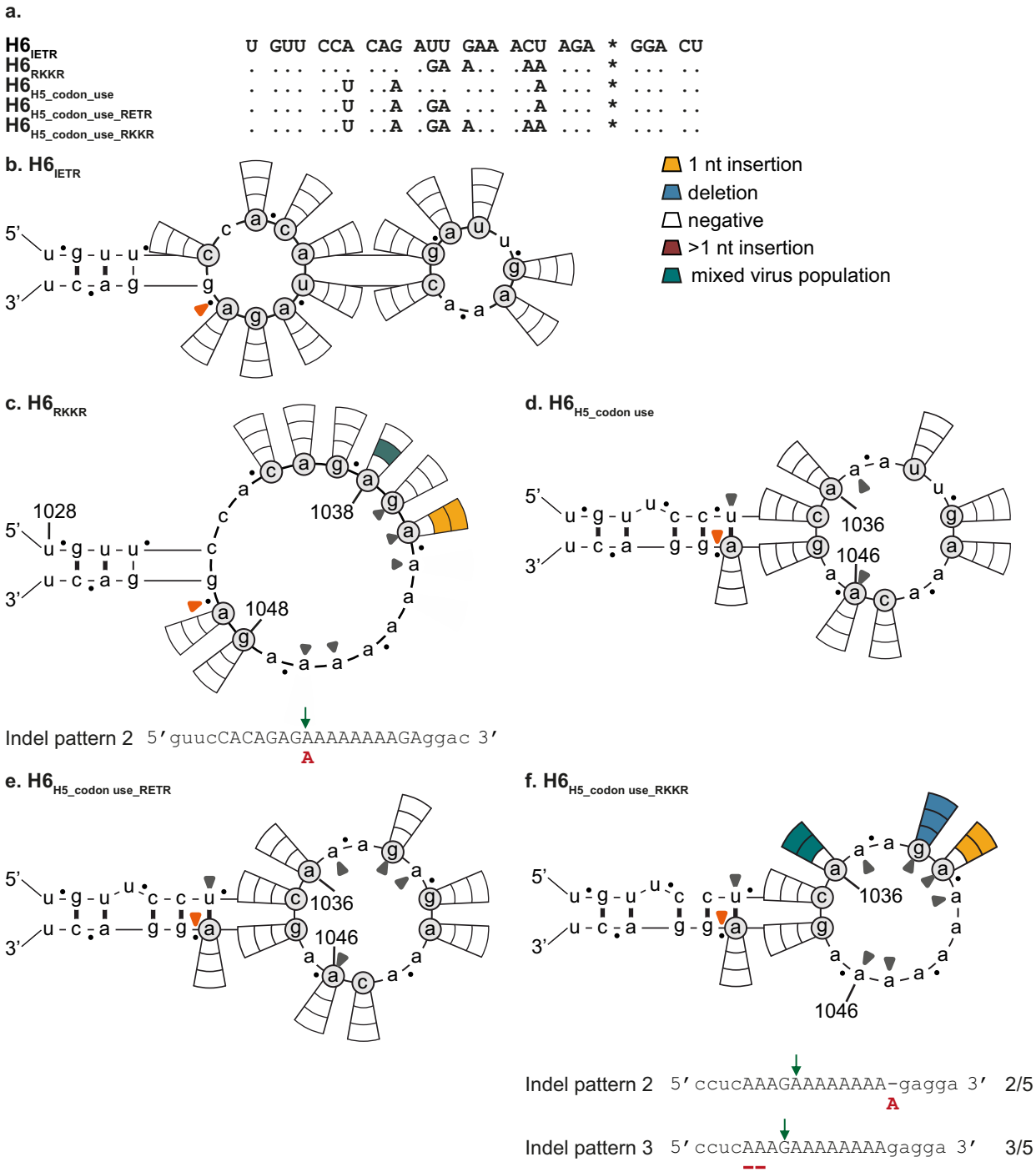


Fig. 3 | Indels are not readily detected at the H6 CS. Results are shown as described in the legend of Fig. 1. Grey closed arrow heads refer to the introduced nucleotide substitutions as compared to $H6_{IETR}$. Observed indel patterns and corresponding frequencies are indicated below the structures, the other detected indels are indicated in Supplementary Table 1a. **a** Alignment of H6 CS sequences shown in Fig. 3. The asterisk indicates the border between HA1 and HA2. Results from testing (**b**) $H6_{IETR}$ HA_{SND}s, (**c**) $H6_{RKKR}$ HA_{SND}s, (**d**) $H6_{H5\_codon\ use}$ HA_{SND}s, (**e**) $H6_{H5\_codon\ use\_RETR}$ HA_{SND}s and (**f**) $H6_{H5\_codon\ use\_RKKR}$ HA_{SND}s.

HPAIV has so far evolved from an H6 LPAIV. SNDs were also introduced into the stem of the H6 CS predicted RNA structure (Supplementary Fig. 1), and no virus was detected in any of the nine experiments (Supplementary Fig. 2c).

Next, we investigated which changes in $H6_{IETR}$ were necessary to detect indels. Firstly, substitutions were introduced to create di- and tribasic intermediate CSs, which consequently increased the number of As at the 3' end of the loop. To this end, the T327K substitution (C1045A and U1046A) was introduced, resulting in an IEKR CS motif and an A-stretch length of six nucleotides. Additionally, a tri-basic cleavage site, shown to be indel-prone in H5, was created by changing U1039G, U1040A, C1045A and U1046A to generate $H6_{REKR}$ (Supplementary Fig. 7a). No indel was observed in any of the $H6_{IEKR}$ and $H6_{REKR}$ HA_{SND}s that were tested (Supplementary Fig. 7b, c), even though a 7-nucleotide long A-stretch was present at the 3'end of the loop of $H6_{IEKR}$ Δ1048, $H6_{REKR}$ Δ1041 and Δ1048 HA_{SND}s. This suggested that, in contrast to H5, a long A-stretch at the 3' end of the loop might not be

sufficient to facilitate indels in H6. Next, an additional G1041A substitution was introduced to create a tetrabasic CS ($H6_{RKKR}$) (Fig. 3a), thereby increasing the length of the A-stretch to eight nucleotides. Indels were then observed in 13% (3/24) of the experiments (Fig. 3c).

As indels were either absent or occurred at a low frequency when H6 HA_{SND}s were tested, we mutated the CS sequence of $H6_{IETR}$ to resemble that of H5. In addition to a different consensus CS sequence at the amino acid level, PQRETR in H5 versus PQIETR in H6, the consensus codon use of the common amino acids differed between the two subtypes[25]. Therefore, three silent substitutions (A1034U, G1037A and U1046A) were first introduced in the A/ml/Sw/81/02 $H6_{IETR}$ to change the P, Q and T codons to those predominantly observed in H5s from viruses of the African-Eurasian-Oceanian lineage ($H6_{H5\_codon\ use}$) (Fig. 3a). No indel was detected in any of the $H6_{H5\_codon\ use}$ HA_{SND}s (Fig. 3d). Therefore, two additional non-silent substitutions (U1039G and U1040A) were introduced in $H6_{H5\_codon\ use}$, resulting in a loop sequence identical to that of $H5_{RETR}$ at both the nucleotide and amino acid levels ($H6_{H5\_codon\ use\_RETR}$). Again, no virus was detected in any of the experiments using $H6_{H5\_codon\ use\_RETR}$ HA_{SND}s (Fig. 3e). Next, additional non-silent substitutions (G1041A and C1045A) were introduced to change the sequence of the $H6_{H5\_codon\ use\_RETR}$ loop to that of $H5_{RKKR}$ ($H6_{H5\_codon\ use\_RKKR}$). Virus was detected in 39% (7/18) of experiments (Fig. 3f), a higher frequency than that observed for $H6_{RKKR}$ (13%). The indels that repaired the HA reading frame corresponded to pattern 2 and 3, which were also observed multiple times in H5. The CS sequence of the virus obtained using $H6_{H5\_codon\ use\_RKKR}$ HA_{SND} Δ1036 could not be reliably detected by Sanger sequencing and sequencing of a subset of TOPO clones revealed a mixed virus population, as observed for $H5_{RKKR}$ (Supplementary Table 1b). As a high indel frequency was observed in $H5_{MBCS}$, the nucleotides that encode the A/Indo/5/05 H5 MBCS (AGA AGA AAA AAG; RRKK) were introduced into $H6_{H5\_codon\ use}$ between T327-R328 ($H6_{H5\_codon\ use\_MBCS}$) (Fig. 4a). Virus was detected in only 11% (4/36) of the experiments (Fig. 4b). Additional substitutions (U1039G, U1040A, A1043G, C1045G and A1046C) were introduced in $H6_{H5\_codon\ use\_MBCS}$ to change the loop sequence completely to that of $H5_{MBCS}$ ($H6_{H5\_codon\ use\_MBCS\_loop\_seq}$) (Fig. 4a). This increased virus detection frequencies from 11 to 44% (20/45; Fig. 4c). Most of the indels observed in $H6_{H5\_codon\ use\_MBCS}$ and $H6_{H5\_codon\ use\_MBCS\_loop\_seq}$ HA_{SND}s corresponded to the previously identified patterns 1 and 2, which were also observed frequently in $H5_{MBCS}$.

Upon increasing the length of the A-stretch at the 3' end of the loop and inserting an H5 MBCS into the H6 CS, indel frequencies remained low. In addition, indel frequency in H6 HA_{SND}s containing the RKKR CS was higher in the presence of the H5 codon use substitutions (13 versus 39%). These HAs only differed by two substitutions, and we sought to investigate the impact of the substitution closest to the CS (G1037A) (Fig. 4a). Introducing G1037A into $H6_{RKKR}$ ($H6_{RKKR\_G1037A}$) HA_{SND}s increased indel frequency from 13 to 61% (11/18; Fig. 4d), which was also higher than observed in the $H6_{H5\_codon\ use\_RKKR}$ HA_{SND}s (39%). Of note, 1037A, although not present in the H6 LPAIV CS consensus, is found in 37.4% of H6 viruses[25]. Indels identified by Sanger sequencing in $H6_{RKKR\_G1073A}$ HA_{SND}s corresponded to indel pattern 2. Nevertheless, as previously observed for $H5_{RKKR}$, $H6_{RKKR}$ and $H6_{H5\_codon\ use\_RKKR}$, the CS sequence of most viruses could not be reliably determined by Sanger sequencing, and a mixed virus population was revealed upon sequencing of TOPO-clones (Supplementary Table 1b). The majority (75%; 33/44) of the sequences showed a frameshift in HA, partially due to the presence of the introduced SND (39%; 13/33). The in-frame insertions mostly led to an additional K in the CS, and a ten A insertion, creating a PQRKKKKKKG CS, was observed.

No direct relationship between indel detection rate and particle production efficiency was observed

In order to investigate whether differences observed in indel frequency could be due to differences in particle production upon reverse genetics, plaque forming units per ml (PFU/ml) were measured. Madin-Darby Canine Kidney (MDCK) cells were inoculated with 293 T supernatants derived from reverse genetics experiments with all HA plasmids used as templates to produce HA_{SND}s, to the exception of those coding for HAs with silent substitutions. Virus titres of $H5_{MBCS}$ or $H5_{RKKR}$ viruses were higher than those of viruses with dibasic ($H5_{RETR}$) and tribasic ($H5_{REKR}$ and $H5_{RKTR}$) CSs, probably due to cleavage independent of trypsin allowing re-amplification in 293T cells (Supplementary Fig. 8). Therefore, the increased indel detection in $H5_{MBCS}$ and $H5_{RKKR}$ HA_{SND}s could be partially due to a higher rescue efficiency and the opportunity for additional rounds of replication. On the other hand, virus titres of $H5_{REKR}$ and $H5_{RKTR}$ viruses were comparable to that of the $H5_{RETR}$ virus even though indels were detected more frequently in $H5_{REKR}$ and $H5_{RKTR}$ than in $H5_{RETR}$ HA_{SND}s.

In general, virus titres of H6 viruses were lower than those of H5 viruses. The number of particles produced upon rescue of $H6_{H5\_codon\ use\_MBCS}$ and $H6_{H5\_codon\ use\_MBCS\_loop\_seq}$ viruses was slightly lower than for the $H5_{MBCS}$ virus, perhaps due to lower cleavage efficiency resulting in lower re-amplification levels. Nevertheless, they were higher than those of the $H5_{REKR}$ and $H5_{RKTR}$ viruses, for which high indel frequencies were observed. Virus titres of $H6_{RKKR}$, $H6_{H5\_codon\ use\_RKKR}$ and $H6_{RKKR\_G1037A}$ viruses were similar or lower than that of the $H6_{IETR}$ virus, yet increased indel detection was observed in the formers. The lowest virus titre was observed for the $H6_{REKR}$ virus, and thus it cannot be excluded that the absence of indels in $H6_{REKR}$ was partially the result of low virus rescue efficiency. Taken together, these results show that no direct relationship between indel detection rate and virus titre was observed. However, it cannot be fully excluded that, for some HAs, indels were not observed because of low virus production and insufficient rounds of genome replication by the RdRp.

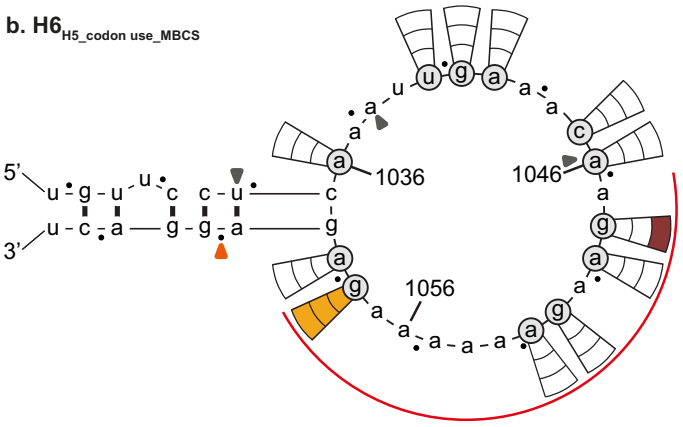
Trans-complementation with a functional H5 protein increased indel detection in the H5 CS but not in that of H6

As a low number of indels were detected using $H5_{RETR}$ HA_{SND}s and no indel was detected with many H6 HA_{SND}s, a *trans*-complementation experiment was performed to attempt to increase indel detection sensitivity. To this end, an $H5_{RETR}$ expression plasmid was co-transfected in 293T cells along with the eight reverse genetics plasmids. This will result in the expression of a functional HA protein, to ensure that infectious virus particles harbouring HA_{SND} genes are produced in 293T cells. An additional round of replication in MDCK cells can then take place, thus increasing the opportunity for indel generation by the RdRp. Addition of the $H5_{RETR}$ expression plasmid during transfection of the $H5_{RETR}$ HA_{SND}s led to an increase in the number of indels that were detected (Fig. 1b and Supplementary Fig. 9a). *Trans*-complementation decreased the clonality of the rescued viruses, whose CS sequences could not be accurately determined using Sanger sequencing. Therefore, six TOPO-clones of HA PCR fragments amplified from MDCK supernatants with an HA titre of ≥8 hemagglutination units (HAU/25 µl) were sequenced. An HA-titre of 8 HAU/25 µl was chosen as it was the lowest titre which coincided with the observation of cytopathic effect in MDCK cells, suggesting virus replication rather than just efficient *trans*-complementation of deficient viruses with an HA_{SND}. TOPO-clones contained different sequences, indeed indicative of the presence of mixed virus populations (Supplementary Table 1b). In total, a positive HA titre of >8 HAU/25 µl was observed with 19/30 (63%) HA_{SND}s, while this was only 13% (4/30) without *trans*-complementation. Frameshift was observed in 11% (13/119) of the clones, which was mostly due to the presence of the introduced SND. In some occasions, the WT sequence was detected, which could be due to $H5_{RETR}$ expression plasmid DNA detection, despite extensive DNAse treatment and negative -RT controls. In total, for seven samples, either only the SND, only the WT, or a combination of both WT and SND were detected in all 6 clones. For one sample, TOPO-cloning failed and thus no sequence was obtained. Other one-nucleotide insertions and deletions were also detected. Taken together, indel detection of $H5_{RETR}$ HA_{SND}s increased upon *trans*-complementation. A similar approach was taken using all H6s in which no indel had been detected ($H6_{IETR}$, $H6_{IEKR}$, $H6_{REKR}$, $H6_{H5\_codon\ use}$ and $H6_{H5\_codon\ use\_RETR}$ HA_{SND}s). An HA-titre of ≥8 HAU/25 µl was detected only in 5% (2/42) and

a.

$H6_{IETR}$	U	GUU	CCA	CAG	AUU	GAA	ACU	---	---	---	---	AGA	*	GGA	CU
$H6_{H5\_codon\_use\_MBCS}$	.	...	..U	..A	...	...	..A	AGA	AGA	AAA	AAG	...	*	...	..
$H6_{H5\_codon\_use\_MBCS\_loop\_seq}$	.	...	..U	..A	.GA	..G	.GC	AGA	AGA	AAA	AAG	...	*	...	..
$H6_{RKKR\_G1037A}$	.	...	...	..A	.GA	A..	.AA	---	---	---	---	...	*	...	..

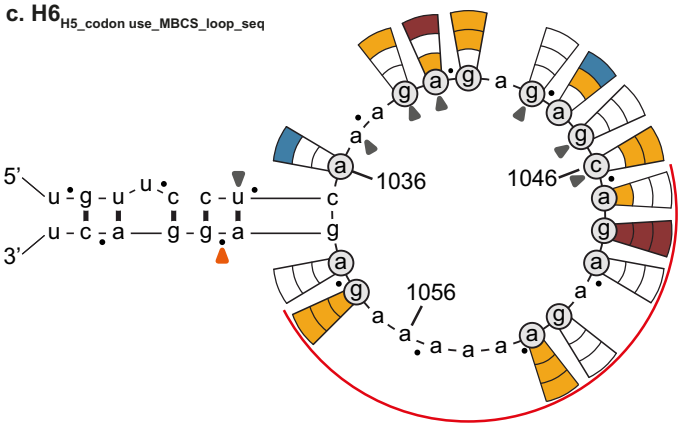
b. $H6_{H5\_codon\ use\_MBCS}$



Indel pattern 1 5′ccucAAAUUGAAACAAGAAGAAAA-------AAGAgagga 3′ 1/4
AAGAAAA

Indel pattern 2 5′ccucAAAUUGAAACAAGAAGAAAAAAGA-gagga 3′ 3/4
A

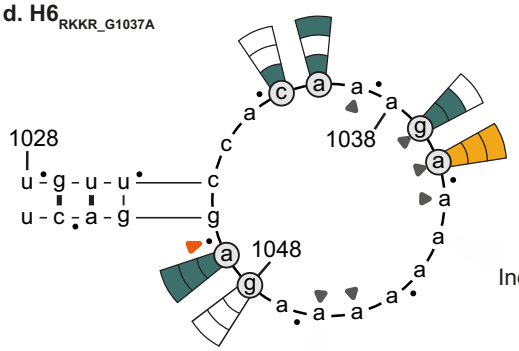
c. $H6_{H5\_codon\ use\_MBCS\_loop\_seq}$



Indel pattern 1 5′ccucAAAGAGAGAGCAGAAGAAAA-------AAGAgagga 3′ 4/20
AAGAAAA

Indel pattern 2 5′ccucAAAGAGAGAGCAGAAGAAAAAA-GAgagga 3′ 14/20
A

d. $H6_{RKKR\_G1037A}$



Indel pattern 2 5′guucCACAAAGAAAAAAAAGAggac 3′ 3/11
A

Fig. 4 | Introducing H5 loop sequences in H6 increases indel detection. Results are shown as described in the legend of Fig. 1. Grey closed arrow heads refer to the introduced nucleotide substitutions as compared to $H6_{IETR}$. Observed indel patterns and corresponding frequencies are indicated below the structures, the other detected indels are shown in Supplementary Table 1a. **a** Alignment of all H6 HA CS sequences shown in Fig. 4. The asterisk indicates the border between HA1 and HA2. Results from testing (**b**) $H6_{H5\_codon\ use\_MBCS}$ HA_{SND}s, (**c**) $H6_{H5\_codon\ use\_MBCS\_loop\_seq}$ HA_{SND}s and (**d**) $H6_{RKKR\_G1037A}$ HA_{SND}s. In (**b**, **c**), the red curved line indicates the location of the inserted MBCS.

8% (1/12) of experiments with $H6_{IETR}$ and $H6_{H5\_codon\ use\_RETR}$ respectively (Supplementary Fig. 9b−f) and only the introduced SND was observed upon Sanger sequencing of the corresponding clones, probably due to very efficient *trans*-complementation. Taken together, these data show that *trans*-complementation increased indel detection in H5, but not in H6.

Discussion

To date, the exact molecular mechanisms underlying MBCS acquisition and the reason(s) why HPAIVs have only evolved from LPAIVs of the H5 and H7 subtypes remain unknown. As reproducing MBCS acquisition in viruses in a laboratory setting has been proven very difficult, we developed a novel sensitive, virus-based system to study the drivers of insertions at the HA CS. Using this system, indels were detected relatively frequently. We observed that A-stretches at both the 5' and 3' end of the loop were the main regions where indels occurred. In addition, four indel patterns were observed, present in 54% (121/226) of the indel-containing viruses. At the 3' end of the loop, 77% (98/127) of indels were insertions, suggesting that insertions in this part of the loop are well tolerated at the RNA and protein levels. At the 5' end of the loop, mostly one-nucleotide insertions (11/37) or two-nucleotide deletions (12/37) were observed, suggesting that larger insertions at this location might be detrimental at the RNA or protein level. Accordingly, reducing the length of the A-stretches in the $H5_{MBCS}$ loop resulted in an increased deletion to insertion ratio that occurred at the 5' end of the loop. Of note, most of the insertions of more than one nucleotide contained one G (32/36) among six (24/36) or three (8/36) As. A-stretches interrupted by Gs might be more prone to duplication through backtracking and re-alignment than A-stretches interrupted by pyrimidines, as Us in the product can pair with both As or Gs in the template during realigning, as previously suggested by Perdue et al.[18]. A six-nucleotide direct repeat consisting of five As and one G has also been observed in previous studies, which either occurred in nature or after in ovo passaging[11,18,39].

Data from the present study show that nucleotide sequence is an important driver of indels. We observed that increasing the A content in $H5_{RETR}$, even through only one nucleotide substitution, increased indel frequency. Of note, it remains unknown whether MBCSs are generated during cRNA and/or vRNA replication and therefore if A and/or U content is a driver of insertions. Our data are in accordance with that of previous studies[20,22,24,26,34,36]. It was observed in ovo, in vivo and in vitro that H5 viruses with an intermediate CS genotype and higher A content, i.e., REKR, RKTR or RKKR, acquired insertions at the CS more easily than viruses with the consensus RETR CS[22,24,26,34,36]. The importance of polypurine-rich regions in insertion acquisition at the H5 CS region has also been proposed by Perdue et al.[18]. Using our experimental system, a trend towards fewer indels was observed upon reducing the length of the A-stretch, even when interrupting it with Gs, suggesting that nucleotide nature (A,C,G,U) rather than base type (purine/pyrimidine) might be a determinant.

This study revealed novel insights regarding the subtype-restriction of MBCS acquisition. Indel detection in H5 increased upon a single nucleotide substitution, when the H5 CS was changed from RETR to REKR or RKTR. In contrast, indels in H6 were only observed when five or more nucleotide substitutions were introduced in the H6 CS. Of note, our experimental system is not fully quantitative, and it cannot be completely excluded that differences in rescue efficiency might lead to small variations in indel detection sensitivity between H5 and H6. Data from the present study are in line with observations from an in-silico analysis that we recently conducted using a comprehensive dataset of all available LPAIV HA CS sequences from all 16 HA subtypes[25]. It was observed that fewer substitutions were required in H5 and H7 HA CS sequences to obtain an insertion-prone sequence, as defined by the presence of three basic amino acids or A-stretches, than in those from other subtypes, including H6. Collectively, these results point to the fact that several substitutions are necessary to generate indels in H6 as a potential limitation for MBCS acquisition in non-H5/H7 subtypes. In addition, while nucleotide sequences of intermediate CS used in this study, e.g., $H5_{REKR}$ and $H5_{RKTR}$, have been detected in WT virus isolates at low frequencies, none of the H6 sequences in which indels were detected in the present study have been observed in natural isolates[25]. Some non-H5/H7 viruses with tribasic CSs have been isolated (e.g., H4N2 and H9N2)[23,40–43], yet they remained trypsin-dependent in vitro[40,41]. The tribasic CS in an H4 virus was acquired by a substitution and insertion, and the tribasic CSs in H9 viruses were acquired by substitutions only. Previously, Zhang et al. showed that changing the H9N2 HA CS to a tri- or tetrabasic CS, thereby increasing the number of consecutive As, increased the detection of insertions as indirectly measured by luciferase activity in a mini-genome assay[23], corroborating that the nucleotide sequence of the CS is a key determinant of insertions. Nevertheless, indel detection in H6 was never as high as in H5 in our study, even when the full $H5_{RKKR}$ or $H5_{MBCS}$ sequence was introduced in H6. Although it cannot be fully ruled out that this could be partially due to a lower particle production in H6 compared to H5, this observation suggests that RNA regions outside of the CS and other factors than CS nucleotide sequence, e.g., RNA structures, might play a role in MBCS acquisition. Several conserved alternative configurations of putative RNA stem-loop structures that encompass CS codons in H5 and H7 HA sequences have been predicted using folding algorithms on naked protein-free RNA, supported by covariation and phylogenetic analyses[19,21,44]. Although it has been proposed that RNA folding in the CS region might drive MBCS acquisition[11,18,22–24,45], strong empirical evidence is lacking. Reducing the loop size of these predicted RNA CS structures, yet retaining a large A-stretch of 12 or nine nucleotides in H5 or H9 HA CSs respectively, reduced indels as indirectly measured by a luciferase-based minigenome assay[22,23]. Kida et al. reported a reduced indel frequency when the length of the A-stretch and the size of the predicted loop were decreased simultaneously, precluding the discrimination between the impact of these two factors on indels[24]. Here, we refrained from investigating the role on MBCS acquisition of these putative RNA structures in MBCS acquisition. These structures are predicted in-silico on naked RNA and might reflect those present in quiescent vRNPs. Instead, it might be more relevant to analyse structures in the context of RNA replication by the influenza RdRp, as RNA structures are expected to be melted by the polymerase complex during replication and might also be altered in the presence of the nucleoprotein.

One limitation of the presently used system is that only frame-restoring indels (insertion of 3n + 1 or deletion of 3n −1 nucleotides) resulting in a functional HA protein are detected. Consequently, the lack of detection of replication competent viruses does not necessarily mean that indels did not occur at all. Other indels not leading to reading frame repair or leading to sequences detrimental to HA structure, folding, expression or function might have occurred but would not be detected using this experimental approach. The development of very sensitive, controlled and reliable next-generation sequencing methods and analyses that allow for the accurate detection of all indels, especially in homopolymer regions, is crucial. Unfortunately, the most widely used next-generation sequencing methods are still unreliable when it comes to the detection of indels in homopolymer regions, which are error-prone for many replication enzymes[46]. Another potential limitation is that our experimental system, based on reverse genetics, relies on the human RNA polymerase I (Pol I) for initial

transcription of viral genomic RNA from the transfected plasmids[47]. Therefore, it cannot be fully excluded that some of the observed indels might have occurred during plasmid transcription. The insertion and deletion rates of Pol I from yeast, estimated at 8.8×10^{-7} and 3.4×10^{-7} per base pair respectively, have been recently investigated using circular sequencing[48]. It was observed that the majority of insertions were one or two nucleotides in length and that, as expected, homonucleotide and dinucleotide tracts were hotspots for indels by the yeast RNA polymerase II, as it is the case for many replication enzymes. Nevertheless, it seems unlikely that Pol I errors majorly contribute to the present data as: (i) indels differed in pattern and length including multiple insertions of four or more nucleotides (30%; 40/133), (ii) no indel was detected in many HA_{SND}s, despite the presence of long homopolymers (notably in H6 HAs and in the homopolymer controls) and (iii) the addition of an MBCS, and thus an A-stretch, in H6 only marginally increased indels, while changing the 5' end of the loop of $H6_{H5\_codon\_use\_MBCS\_loop\_seq}$ to that of H5, without the addition of homopolymers, increased indels.

We have here developed a novel and sensitive experimental setup to study drivers of MBCS acquisition and showed that nucleotide sequence is a key determinant driving insertions at the H5 CS. We have shown for the first time that insertions can be detected at the H6 CS when substitutions increasing the A/U content were introduced. This experimental approach could be further used to identify drivers of indels at the CS of HAs of viruses from other subtypes than H5 and H6. Moreover, the impact of nucleotide substitutions within the CS region could be investigated to identify the minimal number of nucleotide substitutions needed for non-H5 and -H7 viruses to acquire a MBCS. This information could then be used in surveillance programmes to flag LPAIV with potential to evolve to HPAIV.

Methods

Cells

MDCK cells were cultured at 37 °C and 5% CO_2 in EMEM (Capricorn) supplemented with 1.5 mg/ml sodium bicarbonate (Gibco), 10 mM HEPES (Capricorn), 100 IU/ml penicillin (Capricorn), 100 µg/µl streptomycin (Capricorn), 2 mM glutamine (Capricorn), 1X non-essential amino acids (Capricorn) and 10% fetal calf serum (FCS). Human epithelial 293T cells were cultured at 37 °C and 5% CO_2 in DMEM (Capricorn) and supplemented with 1 mM sodium pyruvate (Gibco), 100 Iµ/ml penicillin, 100 µg/µl streptomycin, 2 mM glutamine, 1X non-essential amino acids and 10% FCS. During basal cell culture 293T cells were maintained in the presence of 500 µg/ml geneticin (Gibco) and passaged when sub confluent.

Plasmids

The generation of reverse genetics plasmids (modified version of pHW2000) containing all gene segments of A/Indo/5/05 and the HA of A/ml/Sw/81/02 (accession number: MN515099.1) was described previously[49,50]. The A/Indo/5/05 virus was kindly provided by M. Peiris (Hongkong University). The MBCS of the A/Indo/5/05 HA was removed ($H5_{RETR}$) as described[50]. The generation of reverse genetics plasmids coding for the HA and NA segments of A/PR/8/1934 (H1N1), the HA and M segments of A/Netherlands/602/2009 (H1N1), the HA segment of A/WSN/1933 (H1N1) and the HA segment of A/Guangzhou/39715/2014 (H5N6) were described previously[47,51–53]. The HA gene segment of the H5 A/Indo/5/05 ΔMBCS ($H5_{RETR}$) was cloned into the pCAGGS expression vector, which was kindly provided by Dr. A. Garcia-Sastre (Icahn School of Medicine, New York, U.S.A.).

The plasmids containing the $H5_{RETR}$, H5 A/Indo/5/05 ($H5_{MBCS}$), and H6 A/ml/Sw/81/02 ($H6_{IETR}$) HA were used as templates for site-directed mutagenesis using PFU Ultra II (Agilent Technologies) as described previously[54]. All HA plasmids were confirmed to lead to the expression of functional HA proteins and viruses when used in reverse genetics experiments and were subsequently used as templates to introduce SNDs. Every nucleotide belonging to the H5 LPAIV CS (Q322-R326), H5 HPAIV CS (Q322-R330), H6 CS (Q340-R344) or $H6_{H5\_codon\_use\_MBCS}$ CS (nucleotide 1036-1059) resulting in a unique sequence, were deleted in the HAs described in the main text. A subset of SNDs were introduced into the H5 HAs containing reduced A-stretch lengths. All introduced nucleotide substitutions are indicated in the figures by grey arrowheads and sequences of all the CSs used to introduce SNDs can be viewed in Supplementary Table 1a, c. The presence of the desired mutation in each plasmid preparation was confirmed by Sanger sequencing using a 3130XL or 3500XL genetic analyser (Applied Biosystems). Respective H5 or H6 numberings, starting from the signal peptide, are used throughout the manuscript. Primer sequences are available upon request.

Recombinant virus production and sequencing

Recombinant viruses were produced essentially as described previously[51]. In brief, 293T cells were transfected with 5 µg of a reverse genetics plasmid coding for WT HA or HA_{SND}s, together with 5 µg each of the remaining H5N1 A/Indo/5/05 reverse genetics plasmids (H5 virus backbone). Three days after transfection, undiluted 293T supernatant was used to inoculate MDCK cells and virus production was assessed after three days by hemagglutination assay with 1% turkey red blood cells. If virus was detected, RNA was extracted from 200 µl of the supernatant with the high pure RNA isolation kit (Roche), according to the instructions of the manufacturer. Complementary DNA (cDNA) was produced by reverse transcription (RT) using Superscript IV (Invitrogen) and an influenza A virus specific primer (5'-*AGCRAAAGCAGG*) according to the instructions of the manufacturer. The region in HA spanning the HA CS was amplified by CS PCR using PFU Ultra II and primers 5'-*GGCGATAAACTCTAGTATGC-3*' and 5'-*CGGATAGTTGTACGTTCCGT-3*' for H5, and 5'-*GGTAACAAAAGCTTGCCCTT-3*' and 5'-*ATTCGTGGTCGACAGCTTCG-3*' for H6. PCR products were visualised by gel electrophoresis and agarose bands containing DNA were extracted and purified using the MinElute gel extraction kit (Qiagen) according to the instructions of the manufacturer. Next, sequences of PCR products were obtained by Sanger sequencing using a 3130XL or 3500XL genetic analyser. If ambiguous sequences were obtained by Sanger sequencing, thereby suggesting a mixed virus population, the purified HA band resulting from the H5- or H6-specific CS PCR was cloned using the Zero Blunt TOPO PCR cloning kit (Invitrogen) according to the instructions of the manufacturer. For each sample, 6 clones were sequenced using Sanger sequencing. Samples were considered negative when the Sanger sequencing of all TOPO-clones revealed the presence of the introduced SND, except for the *trans*-complementation samples.

Trans-complementation experiments

Trans-complementation assays were performed as described above with the additional co-transfection of 5 µg of a pCAGGS expression plasmid coding for the $H5_{RETR}$ along with the eight reverse genetics plasmids. In order to deplete plasmid DNA, the MDCK supernatant was passed through a 0.45 µm filter to remove cell debris and the isolated RNA was subjected to extensive DNAse treatment. First, the DNAse I treatment from the high pure RNA isolation kit (Roche) was extended to 30 min. Subsequently, 1 µl of DpnI was added to the extracted RNA followed by incubation for 1 h at 37 °C and DpnI inactivation for 20 min at 80 °C. Turbo DNAse treatment and bead inactivation (Invitrogen) were subsequently performed according to the instructions of the manufacturer. Next, cDNA and HA amplicons were produced as described above. A minus RT control, in which 3 µl of RNA was directly added to the PCR mix, was included in the subsequent HA PCR to confirm the complete removal of plasmid DNA.

Plaque assays

Plaque assays were performed to determine the particle production efficiency upon reverse genetics in 293T cells with most H5 and H6 HAs that were used as templates for the introduction of SNDs. For the mutants containing reduced A-stretch lengths, only the constructs containing non-silent substitutions were tested. The plaque assay was essentially performed

as described previously[54]. In brief, different dilutions of 293T supernatant were added to 6-well plates with confluent MDCK cells. After one hour of incubation at 37 °C and 5% CO_2, the inoculum was removed and cells were washed with PBS twice. Subsequently, 4 ml of a 1:1 dilution of Avicel RC-591 (IMCD) in 2x EMEM (Capricorn) with N-tosyl-L-phenylalanine chloromethyl ketone treated trypsin (Sigma) was added to the plates. Twenty-eight hours after inoculation, the cells were washed twice with PBS and subsequently fixed with 80% acetone. NP staining was performed as described previously[54] and the number of plaques were counted using ImageQuant TL colony counting software version 8.2.0.0 (GE Healthcare, Life sciences).

Biosafety

All experiments with H5N1 and H6N1 viruses were performed under biosafety level 3 or 3^+ conditions.

RNA structure predictions

The predicted conserved RNA structures in the H5 and H6 CS region were described previously[19,21]. RNA structures of the modified H5 and H6 CS stem-loop region and of HA_{SND}s were predicted using the UNAFold Web Server (http://www.unafold.org/mfold/applications/rna-folding-form.php) with default settings. All RNA structure predictions are shown in the positive sense orientation and the indicated 5' and 3' ends of the loop are based on the positive sense orientation.

Data availability

All relevant data are provided within the manuscript and its Supporting Information files. Additional data are available from the corresponding author (M.R.) on request.

Received: 26 December 2023; Accepted: 14 March 2024;
Published online: 13 May 2024

References

1. Shaw M. L. & Palese, P. Orthomyxoviridae. In *Fields Virology* (eds Knipe M. K. & Howley, P. M.) (Lippincott Williams and Wilkins, 2013).
2. Fereidouni, S. et al. Genetic characterization of a new candidate hemagglutinin subtype of influenza A viruses. *Emerg. Microbes Infect.* **12**, 2225645 (2023).
3. Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. Evolution and ecology of influenza A viruses. *Microbiol. Rev.* **56**, 152–179 (1992).
4. Lee, D. H., Criado, M. F. & Swayne, D. E. Pathobiological origins and evolutionary history of highly pathogenic avian influenza viruses. *Cold Spring Harb. Perspect. Med.* **11**, a038679 (2021).
5. Abdelwhab, E. M. & Mettenleiter, T. C. Zoonotic animal influenza virus and potential mixing vessel hosts. *Viruses* **15**, 980 (2023).
6. Klenk, H. D., Rott, R. & Orlich, M. Further studies on the activation of influenza virus by proteolytic cleavage of the haemagglutinin. *J. Gen. Virol.* **36**, 151–161 (1977).
7. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28**, 19–38 (2009).
8. Bosch, F. X., Orlich, M., Klenk, H. D. & Rott, R. The structure of the hemagglutinin, a determinant for the pathogenicity of influenza viruses. *Virology* **95**, 197–207 (1979).
9. 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).
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. Garcia, M., Crawford, J. M., Latimer, J. W., Rivera-Cruz, E. & Perdue, M. L. Heterogeneity in the haemagglutinin gene and emergence of the highly pathogenic phenotype among recent H5N2 avian influenza viruses from Mexico. *J. Gen. Virol.* **77**(Pt 7), 1493–1504 (1996).
12. Suarez, D. L. et al. Recombination resulting in virulence shift in avian influenza outbreak, Chile. *Emerg. Infect. Dis.* **10**, 693–699 (2004).
13. De, B. K., Brownlee, G. G., Kendal, A. P. & Shaw, M. W. Complete sequence of a cDNA clone of the hemagglutinin gene of influenza A/Chicken/Scotland/59 (H5N1) virus: comparison with contemporary North American and European strains. *Nucleic Acids Res.* **16**, 4181–4182 (1988).
14. Veits, J. et al. Avian influenza virus hemagglutinins H2, H4, H8, and H14 support a highly pathogenic phenotype. *Proc. Natl. Acad. Sci. USA* **109**, 2579–2584 (2012).
15. Gohrbandt, S. et al. H9 avian influenza reassortant with engineered polybasic cleavage site displays a highly pathogenic phenotype in chicken. *J. Gen. Virol.* **92**, 1843–1853 (2011).
16. Munster, V. J. et al. Insertion of a multibasic cleavage motif into the hemagglutinin of a low-pathogenic avian influenza H6N1 virus induces a highly pathogenic phenotype. *J. Virol.* **84**, 7953–7960 (2010).
17. Schrauwen, E. J. A. et al. Insertion of a multibasic cleavage site in the haemagglutinin of human influenza H3N2 virus does not increase pathogenicity in ferrets. *J. Gen. Virol.* **92**, 1410–1415 (2011).
18. Perdue, M. L., Garcia, M., Senne, D. & Fraire, M. Virulence-associated sequence duplication at the hemagglutinin cleavage site of avian influenza viruses. *Virus Res.* **49**, 173–186 (1997).
19. Gultyaev, A. P., Richard, M., Spronken, M. I., Olsthoorn, R. C. L. & Fouchier, R. A. M. Conserved structural RNA domains in regions coding for cleavage site motifs in hemagglutinin genes of influenza viruses. *Virus Evol.* **5**, vez034 (2019).
20. Abolnik, C. Evolution of H5 highly pathogenic avian influenza: sequence data indicate stepwise changes in the cleavage site. *Arch. Virol.* **162**, 2219–2230 (2017).
21. Gultyaev, A. P. et al. Subtype-specific structural constraints in the evolution of influenza A virus hemagglutinin genes. *Sci. Rep.* **6**, 38892 (2016).
22. Nao, N. et al. Genetic predisposition to acquire a polybasic cleavage site for highly pathogenic avian influenza virus hemagglutinin. *mBio* **8**, e02298–16 (2017).
23. Zhang, J. et al. A risk marker of tribasic hemagglutinin cleavage site in influenza A (H9N2) virus. *Commun. Biol.* **4**, 71 (2021).
24. Kida, Y. et al. Structural requirements in the hemagglutinin cleavage site-coding RNA region for the generation of highly pathogenic avian influenza virus. *Pathogens* **10**, 1597 (2021).
25. Funk, M., de Bruin, A. C. M., Spronken, M. I., Gultyaev, A. P. & Richard, M. In silico analyses of the role of codon usage at the hemagglutinin cleavage site in highly pathogenic avian influenza genesis. *Viruses* **14**, 1352 (2022).
26. Ito, T. et al. Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. *J. Virol.* **75**, 4439–4443 (2001).
27. Ohuchi, M. et al. Mutations at the cleavage site of the hemagglutinin after the pathogenicity of influenza virus A/chick/Penn/83 (H5N2). *Virology* **168**, 274–280 (1989).
28. Orlich, M., Gottwald, H. & Rott, R. Nonhomologous recombination between the hemagglutinin gene and the nucleoprotein gene of an influenza virus. *Virology* **204**, 462–465 (1994).
29. Li, S. Q., Orlich, M. & Rott, R. Generation of seal influenza virus variants pathogenic for chickens, because of hemagglutinin cleavage site changes. *J. Virol.* **64**, 3297–3303 (1990).
30. Orlich, M., Khatchikian, D., Teigler, A. & Rott, R. Structural variation occurring in the hemagglutinin of influenza virus A/turkey/Oregon/71 during adaptation to different cell types. *Virology* **176**, 531–538 (1990).
31. Brugh, M. B. Jr. Recovery of minority subpopulations of highly pathogenic avian influenza virus. *Avian Dis.* **47**, 166–174 (2003).

32. Khatchikian, D., Orlich, M. & Rott, R. Increased viral pathogenicity after insertion of a 28S ribosomal RNA sequence into the haemagglutinin gene of an influenza virus. *Nature* **340**, 156–157 (1989).
33. Laleye, A. T. & Abolnik, C. Emergence of highly pathogenic H5N2 and H7N1 influenza A viruses from low pathogenic precursors by serial passage in ovo. *PLoS One* **15**, e0240290 (2020).
34. Horimoto, T. & Kawaoka, Y. Molecular changes in virulent mutants arising from avirulent avian influenza viruses during replication in 14-day-old embryonated eggs. *Virology* **206**, 755–759 (1995).
35. Seekings, A. H. et al. The emergence of H7N7 highly pathogenic avian influenza virus from low pathogenicity avian influenza virus using an in ovo embryo culture model. *Viruses* **12**, 920 (2020).
36. Luczo, J. M. et al. Evolution of high pathogenicity of H5 avian influenza virus: haemagglutinin cleavage site selection of reverse-genetics mutants during passage in chickens. *Sci. Rep.* **8**, 11518 (2018).
37. Lin, W. et al. Evolution and pathogenicity of H6 avian influenza viruses isolated from Southern China during 2011 to 2017 in mice and chickens. *Sci. Rep.* **10**, 20583 (2020).
38. Everest, H. et al. The evolution, spread and global threat of H6Nx Avian influenza viruses. *Viruses* **12**, 673 (2020).
39. Perdue, M. L., Garcia, M., Beck, J., Brugh, M. & Swayne, D. E. An Arg-Lys insertion at the hemagglutinin cleavage site of an H5N2 avian influenza isolate. *Virus Genes* **12**, 77–84 (1996).
40. Wong, S. S. et al. Characterization of an H4N2 influenza virus from Quails with a multibasic motif in the hemagglutinin cleavage site. *Virology* **468-470**, 72–80 (2014).
41. Parvin, R. et al. Comparison of pathogenicity of subtype H9 avian influenza wild-type viruses from a wide geographic origin expressing mono-, di-, or tri-basic hemagglutinin cleavage sites. *Vet. Res.* **51**, 48 (2020).
42. Begum, J. A. et al. Experimental pathogenicity of H9N2 Avian influenza viruses harboring a Tri-Basic Hemagglutinin Cleavage Site in Sonali and Broiler Chickens. *Viruses* **15**, 461 (2023).
43. Wen, F. et al. Identification of a duck H9N2 influenza virus possessing tri-basic hemagglutinin cleavage sites genetically close to the human H9N2 isolates in China, 2022. *J. Infect.* **86**, e153–e155 (2023).
44. Dupre, G. et al. Phylodynamic study of the conserved RNA structure encompassing the hemagglutinin cleavage site encoding region of H5 and H7 low pathogenic avian influenza viruses. *Virus Evol.* **7**, veab093 (2021).
45. Beerens, N. et al. Emergence and selection of a highly pathogenic avian influenza H7N3 virus. *J. Virol.* **94**, e01818–e01819 (2020).
46. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom. Bioinform.* **3**, lqab019 (2021).
47. Hoffmann, E., Neumann, G., Kawaoka, Y., Hobom, G. & Webster, R. G. A DNA transfection system for generation of influenza A virus from eight plasmids. *Proc. Natl. Acad. Sci. USA* **97**, 6108–6113 (2000).
48. Gout, J. F. et al. The landscape of transcription errors in eukaryotic cells. *Sci. Adv.* **3**, e1701484 (2017).
49. Keawcharoen, J. et al. Repository of Eurasian influenza A virus hemagglutinin and neuraminidase reverse genetics vectors and recombinant viruses. *Vaccine* **28**, 5803–5809 (2010).
50. Chutinimitkul, S. et al. In vitro assessment of attachment pattern and replication efficiency of H5N1 influenza A viruses with altered receptor specificity. *J. Virol.* **84**, 6825–6833 (2010).
51. de Wit, E. et al. Efficient generation and growth of influenza virus A/PR/8/34 from eight cDNA fragments. *Virus Res.* **103**, 155–161 (2004).
52. Herfst, S. et al. Introduction of virulence markers in PB2 of pandemic swine-origin influenza virus does not result in enhanced virulence or transmission. *J. Virol.* **84**, 3752–3758 (2010).
53. Herfst, S. et al. Human Clade 2.3.4.4 A/H5N6 influenza virus lacks mammalian adaptation markers and does not transmit via the airborne route between Ferrets. *mSphere* **3**, e00405–e00417 (2018).
54. Gultyaev, A. P. et al. RNA structural constraints in the evolution of the influenza A virus genome NP segment. *RNA Biol.* **11**, 942–952 (2014).

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation program under DELTA-FLU (grant agreement 727922), NIH/NIAID contract number HHSN272201400008C and NWO ENW M1 grant (OCENW.M.21.150). M.F. was funded through NWO ENW XS grant (OCENW.XS22.1.121) and ZonMW Off Road (04510012010056).

Author contributions

M.S., M.F. and M.R. designed the experiments. M.S. performed the experiments and data analysis. M.S., M.F., A.C.M.dB. and M.R. discussed the data. M.S. and M.R. wrote the initial manuscript. M.F., A.C.M.dB., A.P.G. and R.A.M.F. edited the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44298-024-00029-1.

Correspondence and requests for materials should be addressed to Mathilde Richard.