BMC Genomics

RESEARCH ARTICLE   Open Access

G-quadruplexes in H1N1 influenza genomes

Václav Brázda[1,2*†], Otília Porubiaková[1,2†], Alessio Cantara[1,3], Natália Bohálová[1,3], Jan Coufal[1], Martin Bartas[4], Miroslav Fojta[1] and Jean-Louis Mergny[1*]

Abstract

Background: Influenza viruses are dangerous pathogens. Seventy-Seven genomes of recently emerged genotype 4 reassortant Eurasian avian-like H1N1 virus (*G4*-EA-H1N1) are currently available. We investigated the presence and variation of potential G-quadruplex forming sequences (PQS), which can serve as targets for antiviral treatment.

Results: PQS were identified in all 77 genomes. The total number of PQS in *G4*-EA-H1N1 genomes was 571. Interestingly, the number of PQS per genome in individual close relative viruses varied from 4 to 12. PQS were not randomly distributed in the 8 segments of the *G4*-EA-H1N1 genome, the highest frequency of PQS being found in the NP segment (1.39 per 1000 nt), which is considered a potential target for antiviral therapy. In contrast, no PQS was found in the NS segment. Analyses of variability pointed the importance of some PQS; even if genome variation of influenza virus is extreme, the PQS with the highest G4Hunter score is the most conserved in all tested genomes. G-quadruplex formation in vitro was experimentally confirmed using spectroscopic methods.

Conclusions: The results presented here hint several G-quadruplex-forming sequences in *G4*-EA-H1N1 genomes, that could provide good therapeutic targets.

Keywords: Influenza virus, G-quadruplex, G4Hunter

Background

Influenza viruses are deadly pathogens for humans, and more generally mammals, as well as avian species. They belong to the *Orthomyxoviridae* family and are classified into three types termed Influenza A, B and C. Among these, influenza A viruses (IAVs) pose the greatest threat to human and animal health. IAV genome is divided to 8 segments of negative-sense RNA that encodes 11 proteins [1]. Subtype classification of *G4*-EA-H1N1 is based on the antigenicity of the two major cell surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA). HA protein facilitates binding of the virus to host cell receptors and subsequent endosomal fusion [2], and NA protein is

responsible for binding to cellular receptors and fusion of the viral membranes, causing replication and transcription of viral RNAs in the infected host [3, 4]. The viral RNA genome (gRNA) is transcribed into mRNA and replicated through an intermediate RNA to produce a large quantity of progeny gRNA. These NAs are synthesized by the viral RNA-dependent RNA polymerase complex – polymerase basic protein 2 (PB2), polymerase basic protein 1 (PB1) and polymerase acidic protein (PA), the nucleoprotein (NP), the matrix protein (M) and the non-structural protein (NS) [5, 6].

Roots of virus H1N1 can be traced to 1918, when an avian virus overcame the species barrier to infect humans [7]. That was the beginning of a pandemic that resulted in an estimated 50 to 100 million deaths. Thereafter, influenza viruses rapidly diverged antigenically and three years later this virus was replaced by a new strain. Reassortment of influenza viruses is a major mechanism

---

* Correspondence: vaclav@ibp.cz; mergny@ibp.cz
†Václav Brázda and Otília Porubiaková contributed equally to this work.
[1]Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, 612 65 Brno, Czech Republic
Full list of author information is available at the end of the article

Brázda et al. BMC Genomics (2021) 22:77

Page 2 of 11

to generate progeny viruses with novel antigenic and biological characteristics [8, 9]. The emerged geno-type 4 reassortant Eurasian avian-like H1N1 virus (*G4*-EA-H1N1) has become predominant in swine populations since 2016 [10] and is a new cause of concern.

Guanine quadruplexes (G4) are local nucleic acid structures formed by G-rich DNA and RNA in which four guanines fold in a planar arrangement through Hoogsteen hydrogen bonds [11, 12]. Putative quadruplex sequences (PQSs) contribute to the regulation of key biological processes [13] and have been found in the genomes of viruses (reviewed in: [14]). For example, it has been demonstrated that G-quadruplexes regulate HIV transcription and can be targeted by small compounds called G4 ligands. A comprehensive database of PQS in human all human viruses found with the Quadparser algorithm has been published [15] but these new H1N1 strains were not available at that time.

Here we analyzed 77 newly sequenced variations of H1N1 influenza virus emerged during the last years with a different algorithm, G4Hunter. There are accessible several tools to analyze PQS in genomic sequences (reviewed in [16]). We used the G4Hunter algorithm where G4 propensity is calculated depending on G richness and G/C skewness and PQS are evaluated quantitatively [17] and validated experimentally [17, 18]. We used a new G4Hunter algorithm implementation, which is suitable for batch and full genomes analyses [19, 20] and accessible as the web-tool G4Hunter web [21]. Analyses of the human genome revealed the presence of many G4-prone sequences and G4 presence has been demonstrated in a variety of species, including eukaryotes, bacteria, archaea or viruses both in silico [19, 20, 22] and confirmed experimentally [17, 23, 24]. G4 have been shown to participate in cellular and viral replication, recombination and control of gene expression [25–27]. In addition, DNA aptamers that adopt a quadruplex fold have been described as inhibitors and diagnostic tools to detect viruses [28].

In this article, we analyzed 77 *G4*-EA-H1N1 virus genomes for G-quadruplex occurrence, localization and variance to provide a rational background for PQS targeting in antiviral influenza therapy approaches.

Results

We analyzed 616 sequences in total belonging to 77 strains of *G4*-EA-H1N1. The genome of *G4*-EA-H1N1 is 13,133 nt long and consists of 8 different segments: PB1, PB2, M, HA, NP, NS, PA and NA. PQS frequencies were analyzed according to individual *G4*-EA-H1N1 strains, and for statistical comparison we have grouped genomes according to regions of origin (10 groups based on [10]) and also according to their genomic segments (8

segments). The average GC content for the entire list of viruses is 43.37%, with minimal differences between strains, from 43.20% in the Heilongjiang strain to 43.44% in the Shandong strain. Using standard default values for the G4Hunter algorithm (window size of 25 nucleotides and G4Hunter score above 1.2), 571 PQSs were found among all genomes and all fragments. Mean PQS frequency for the whole set of sequences was 0.56 PQS per 1000 nt and PQSs cover an average of 1.58% of *G4*-EA-H1N1 genomes. The mean number of PQS per *G4*-EA-H1N1 genome was 7.42. The highest number of PQS was found in Swine Beijing 0301 2018 strain with a total of 12 PQSs, giving a PQS frequency of 0.91 PQS per 1000 nt. The lowest frequency (0.30 PQS per 1000 nt) was found in Swine Shandong S113 2014 and Swine Shandong JM78 2017 strains, where only 4 PQS with a G4Hunter score above 1.2 were found. Genomic sequence sizes, GC count, and PQS characteristics are summarized in **Table 1**, all results for individual species and groups are in SM_02A.

Our analyses showed that PQS frequencies of *G4*-EA-H1N1 were significantly different for the Shandong group (compared to Hebei ($p = 0.016$), Jiangsu ($p = 0.047$), Liaoning ($p = 0.0041$) groups), and for the Liaoning group (compared to Henan ($p = 0.025$) and Heilongijang groups ($p = 0,031$)) (available in SM_03). Graphical representation of PQS frequencies is shown in Fig. 1.

We also performed PQS analyses of individual segments of influenza genomes (**Table 2.**); all results for segments are shown in SM_02B. Even if the global GC content in all species is very conserved, the GC content within each segment is more variable - from 41.16% in the HA segment to 47.34% in the M segment. Despite the highest CG content in HA segment, the highest mean PQS frequency was found in the NP segment (with a GC content of 46.23%), with the highest number of PQS (160). It was followed by segments NA (149 PQS) and PB2 (79 PQS). On the other hand, no PQS was found in the NS segment (which codes the non-structural protein) with a GC content of 41.52%. These data are pointing to possible functional importance of G-quadruplex in IAV genomes. All the species have 1, 2 or 4 PQS in segment NP, except for Swine Shandong LY142 2017, which does not contain any PQS with a G4Hunter score above 1.2. IAV belong to the negative-sense single-stranded RNA viruses group. Interestingly, the PQS were not distributed equally among minus gRNA which is copied for protein production (mRNA). Most of the PQSs are located in its mRNA (498 compare to 73 in gRNA). Moreover, in PB1, PB2, NP and NA segments PQS are exclusively found in mRNA (**Table 2**).

The distribution of G4Hunter score parameters for all PQSs found in *G4*-EA-H1N1 segments is summarized in **Table 3**. As previously found in eukaryotes, bacteria and

Table 1 Strains of *G4*-EA-H1N1: genomic sequences sizes, PQS frequency and total counts of PQS. Seq (number of strains), Length (length of the sequence, nt), GC % (average GC content), PQS (total number of predicted PQS), Mean PQS (mean number of predicted PQS), Min PQS (lowest number of predicted PQS), Max PQS (highest number of predicted PQS), PQS frequency (PQS frequency per 1000 nt), Cov% (% of genome covered by PQS).

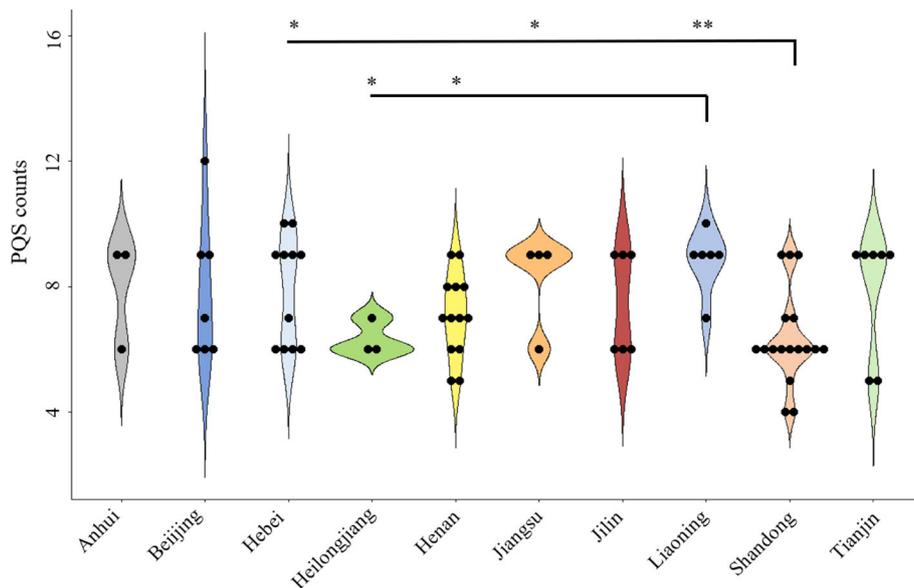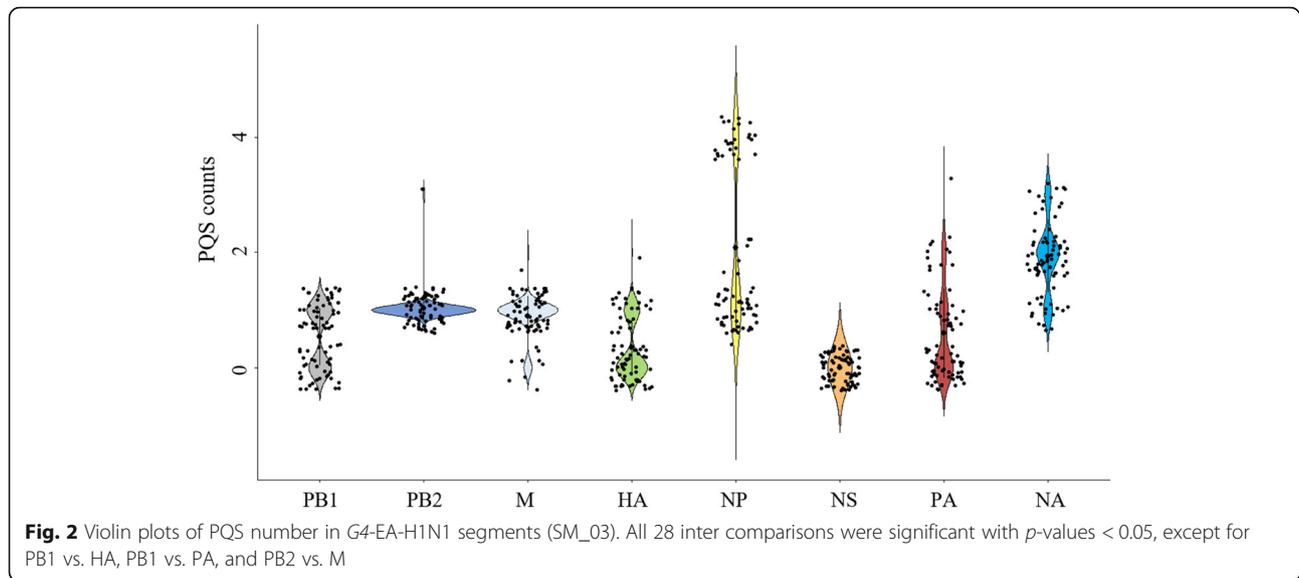|	Seq	Length	GC%	PQS	Mean PQS	Min PQS	Max PQS	PQS frequency	Cov%
All	77	13 133	43.37	571	7.42	4	12	0.56	1.58
Strain	**Seq**	**Length**	**GC%**	**PQS**	**Mean PQS**	**Min PQS**	**Max PQS**	**PQS frequency**	**Cov%**
Anhui	3	13 133	43.25	24	8.00	6	9	0.61	1.64
Beijing	7	13 133	43.39	55	7.86	6	12	0.60	1.70
Hebei	11	13 133	43.35	87	7.91	6	10	0.60	1.68
Heilongjiang	3	13 133	43.20	19	6.33	6	7	0.48	1.30
Henan	13	13 133	43.38	92	7.08	5	9	0.54	1.51
Jiangsu	4	13 133	43.29	33	8.25	6	9	0.63	1.71
Jilin	6	13 133	43.39	45	7.5	6	9	0.57	1.61
Liaoning	6	13 133	43.41	53	8.83	7	10	0.67	1.88
Shandong	17	13 133	43.44	108	6.35	4	9	0.48	1.34
Tianjin	7	13 133	43.30	55	7.86	5	9	0.60	1.70

Fig. 1 Violin plot of PQS number in *G4*-EA-H1N1 groups (SM_03). The significant differences between groups are depicted by asterisks (*p*-value < 0.05 is *; *p*-value < 0.01 is **)

Table 2 Segments of *G4*-EA-H1N1: genomic sequences sizes, PQS frequencies and total counts of PQS. Seq (total number of sequences), Length (median length of sequences), GC % (average GC content), PQS (total number of predicted PQS), Mean PQS (mean number of predicted PQS per sequence), Min PQS (lowest number of predicted PQS per sequence), Max PQS (highest number of predicted PQS per sequence), mRNA-gRNA (viral messenger RNA-viral genome RNA), PQS frequency (PQS frequency per 1000 nt), Cov% (% of genome covered by PQS).

	Seq	Length	GC%	PQS	Mean PQS	Min PQS	Max PQS	PQS frequency	Cov%
All	616	13133	43.37	571	0.93	0	4	0.56	1.58

Segment	Seq	Length	GC%	PQS	Mean PQS	Min PQS	Max PQS (mRNA-gRNA)	PQS frequency	Cov%
PB1	77	2274	42.39	41	0.53	0	1 (1-0)	0.23	0.73
PB2	77	2280	43.51	79	1.03	0	3 (3-0)	0.45	1.23
M	77	982	47.34	69	0.9	0	2 (0-2)	0.91	2.29
HA	77	1701	41.16	27	0.35	0	2 (1-1)	0.21	0.55
NP	77	1497	46.23	160	2.08	0	4 (4-0)	1.39	3.69
NS	77	838	43.36	0	0	0	0	0	0
PA	77	2151	43.42	46	0.60	0	3 (2-1)	0.28	0.71
NA	77	1410	41.52	149	1.94	1	3 (3-0)	1.37	4.25

viruses [19, 20, 22], most of the PQS have relatively low G4Hunter scores (in the 1.2–1.4 range). Only 10 / 571 motifs have a G4Hunter score above 1.4 (all in the HA segment), and no PQS was found with a G4Hunter above 1.6.

Detailed statistical characteristics for PQS frequencies per 1000 nt, including mean, variance, and outliers, are depicted in boxplots for segments are shown in Fig. 2. Statistical evaluation of PQS in IAV segments showed the statistical differences for all comparisons except for

three cases (PB1 *vs.* HA, PB1 *vs.* PA, and PB2 *vs.* M) for which differences were not significantly different.

We evaluated the localization of G4 prone sequences in the genome of Swine Beijing 0301 2018, where we found the highest number of PQS (Fig. 3.). From a total of 12 PQS found, 3 PQSs were in the PB2 and NA segments, 2 PQSs were located in the NP and PA segments and 1 PQS was found in the M and HA segments. The majority of PQS were found in mRNA. Ten out of all PQS were located in mRNA (with positive G4Hunter

Table 3 PQSs in *G4*-EA-H1N1 segments grouped by G4Hunter score (absolute values). Frequency was computed using total number of PQSs in each category divided by total length of all analyzed sequences and multiplied by 1000, the total number of PQS are in brackets.

PQS frequency per 1,000 nt (number of PQS)								
	PB1	PB2	M	HA	NP	NS	PA	NA
1.2-1.4	0.23 (41)	0.45 (79)	0.91 (69)	0.13 (17)	1.39 (160)	0	0.28 (46)	1.37 (149)
1.4-1.6	0	0	0	0.08 (10)	0	0	0	0
>1.6	0	0	0	0	0	0	0	0

Fig. 2 Violin plots of PQS number in *G4*-EA-H1N1 segments (SM_03). All 28 inter comparisons were significant with *p*-values < 0.05, except for PB1 vs. HA, PB1 vs. PA, and PB2 vs. M

score), whereas only 2 PQS were located in negative genomic RNA (with negative G4Hunter score). Interestingly, in segment M, one PQS was located at the 3′ end of intron in negative-sense genomic RNA, near the splicing site of mRNA, which encodes M2 protein. M segment codes 2 matrix proteins – M1, which is coded by whole segment and spliced protein M2 [29]. All 10 conserved PQSs located in positive-sense RNA completely span coding regions; this is hardly surprising, as the vast majority of RNA segments are protein coding, except for short 3′ and 5′ UTRs.

A comparison of genomes revealed that some, but not all, PQS motifs were highly conserved. We align all predicted PQS and generate their LOGO representation (SM_04). Selected LOGO sequences with the highest positive and negative G4Hunter scores and with the most variable nucleotides are shown in Fig. 4. For example, in the M and HA segments, we found PQS in which only 1 nucleotide (out of 25 and 27, respectively) is variable within the PQS motif among all 77 strains. In contrast, other PQS sequences were poorly conserved / extremely variable (for example, the PQS sequence "**C**" in the NP segment has 12 / 26 variable nucleotides in its PQS; this can lead to significant variations in G4Hunter score and quadruplex propensity).

Overall, *G4*-EA-H1N1 genomic sequences are very variable. The analyses of 77 *G4*-EA-H1N1 genomes show a global variation of 23.4%. Therefore, the high sequence conservation of some PQS (two of them have a variation < 4.0% in Fig. 4) suggests they play crucial roles in influenza virus. The PQS sequence with the highest G4Hunter score is also the most conserved among all found PQS. Similarly, another sequence with two GGGG

runs (Fig. 4d), which could form bimolecular G4, has 100% conservation within the G-tracts.

We then determined if the quadruplex-prone sequences identified in silico actually form G4 in vitro. This experimental confirmation is important for these motifs, as their G4-Hunter scores are relatively low, and some candidate sequences may prefer formation of other structures and/or fail to form stable G4 (100% confidence in predicted motifs can only be achieved for relatively high scores, typically above 1.6). To confirm the ability of the most conserved PQS to form G4 in vitro, we used a combination of two biophysical methods, circular dichroism (CD) spectroscopy and the Thioflavin T (ThT) fluorescent assay [30, 31], results are shown in SM_06. We tested nine synthetic oligonucleotides derived from the LOGO sequence listed in Fig. 4. For sequences A, C and E we analyzed two variants, one with the highest and one with the lowest possible G4Hunter score. Quadruplex formation was confirmed for 5 out of 8 analyzed sequences (**Table 4**). G-quadruplex formation in vitro was confirmed by CD spectroscopy as the shift of the peak from 270 to 264 nm and a stronger signal in the presence of K^+ ions (potassium ions stabilize the G4 structure). An example of positive result is presented in Fig. 5, **part A** for a conserved sequence derived from HA fragment and in Fig. 5, **part C** for the sequence from NP fragment with the highest possible G4Hunter score. An example of negative result acquired by CD spectroscopy is shown in Fig. 5, **part B** for a negative control sequence with the G4Hunter score of 0.37 and in Fig. 5, **part D** for the sequence derived from the NP fragment with the lowest possible G4Hunter score.
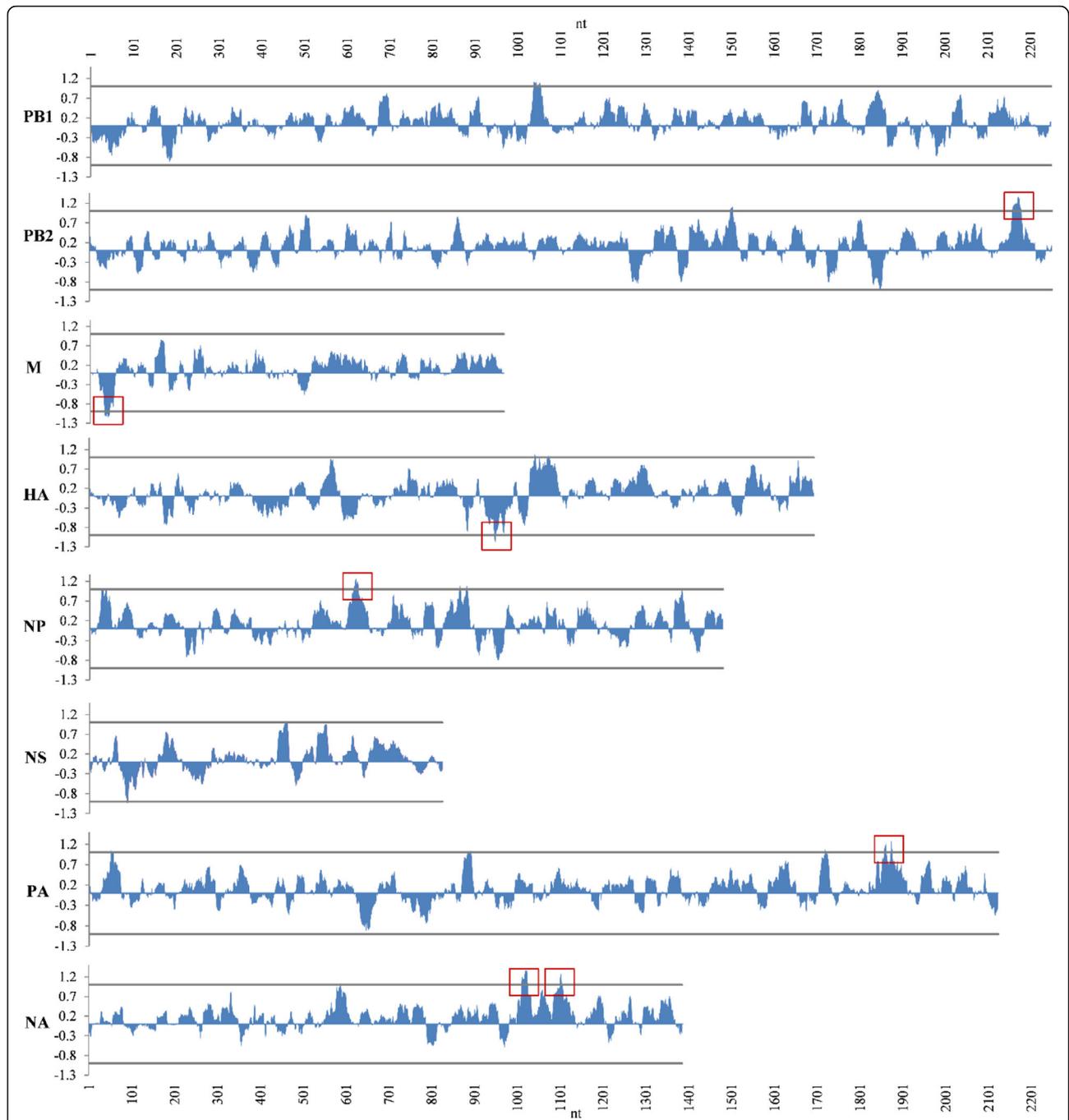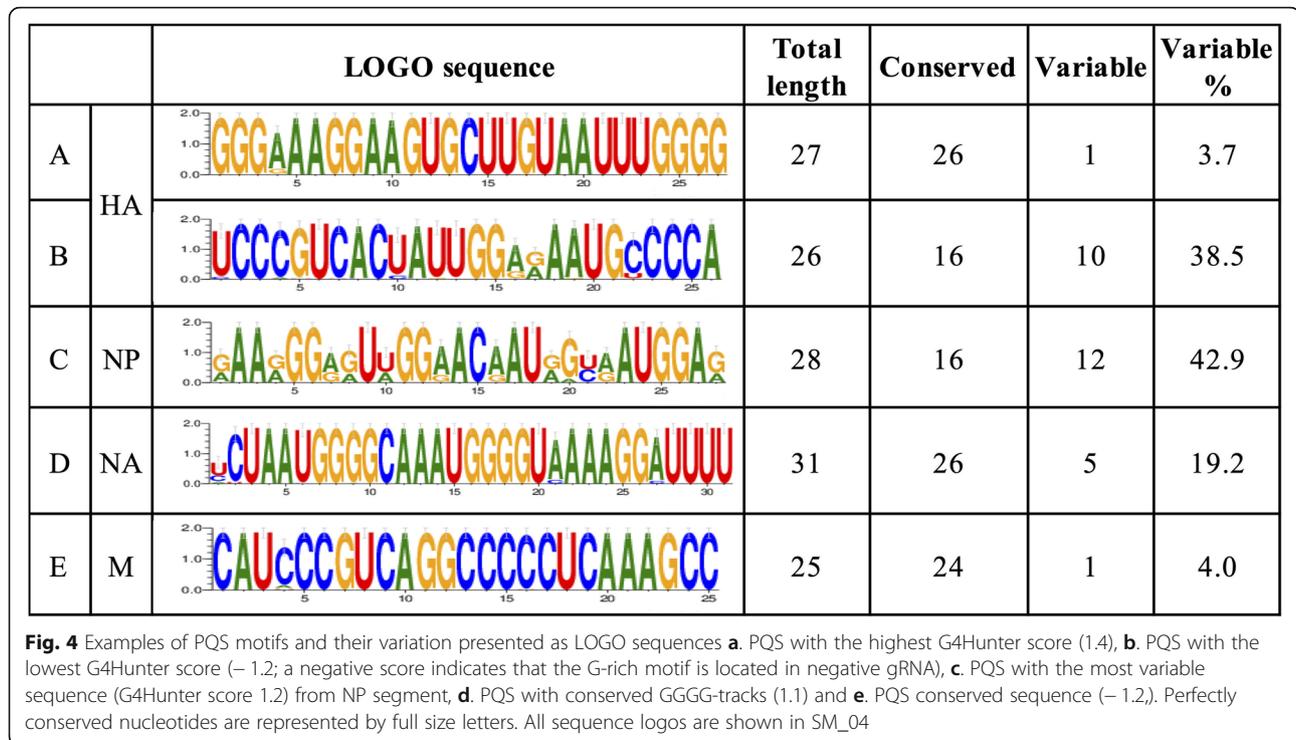
Fig. 3 Localization of G4 prone sequences in the genome of Swine Beijing 0301 2018. Y-axis represents G4Hunter score, x-axis the length of segments. Grey lines define G4Hunter score with value of 1. PQS identified by G4Hunter with G4Hunter score over 1.2 are highlighted by red rectangles

Discussion and conclusions

The influenza viruses pose a global public health concern. Influenza claims 250,000–500,000 lives annually, even though vaccines and antiviral drugs are available. There is therefore an urgent need to develop antiviral drugs with novel mechanisms of action. Noncanonical nucleic acid structures play an important role in basic biological processes [32] and it has been shown that G4s

may be used as targets for therapy [33, 34]. Therefore, noncanonical structures in the H1N1 viral genome could serve as possibly targets for antiviral therapy. In this study, we provide a detailed analysis of PQSs occurrences, frequencies and distributions in the contemporary emerged *G4*-EA-H1N1 strains.

We found a total number of 571 PQS in all 77 *G4*-EA-H1N1 genomes. Interestingly, the number of PQS in
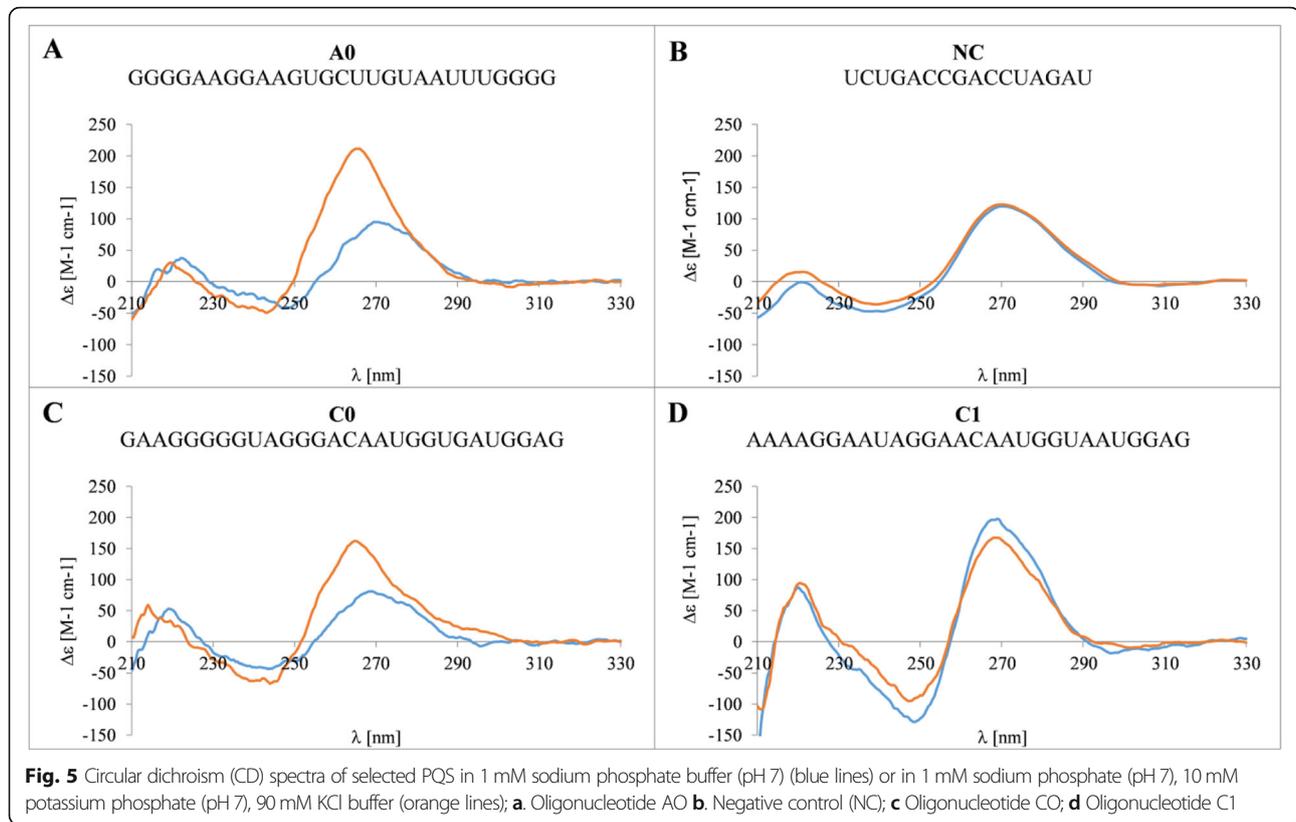
		LOGO sequence	Total length	Conserved	Variable	Variable %
A	HA	GGGAAAGGAAGUGCUUGUAAUUUGGGG	27	26	1	3.7
B		UCCCGUCACUAUUGGAGAAUGCCCCA	26	16	10	38.5
C	NP	GAAGGGAGUUGGAACAAUGGUAAUGGAG	28	16	12	42.9
D	NA	UCUAAUGGGGCAAAUGGGGUAAAAGGAUUUU	31	26	5	19.2
E	M	CAUCCCGUCAGGCCCCCUCAAAGCC	25	24	1	4.0

Fig. 4 Examples of PQS motifs and their variation presented as LOGO sequences **a**. PQS with the highest G4Hunter score (1.4), **b**. PQS with the lowest G4Hunter score (− 1.2; a negative score indicates that the G-rich motif is located in negative gRNA), **c**. PQS with the most variable sequence (G4Hunter score 1.2) from NP segment, **d**. PQS with conserved GGGG-tracks (1.1) and **e**. PQS conserved sequence (− 1.2,). Perfectly conserved nucleotides are represented by full size letters. All sequence logos are shown in SM_04

close *G4*-EA-H1N1 relatives varied from 4 to 12. Analyses of variability pointed to the importance of some PQS: even if genome variation of influenza virus is extreme, the PQS with the highest G4Hunter score is nearly perfectly conserved in all tested genomes. Comparison of segments shows significant differences among individual *G4*-EA-H1N1 segments. While the highest mean PQS frequency was found in the NP segment (1.39), which codes for a protein playing a central role in

viral replication [35], the most abundant viral protein in infected cells [36, 37] and the most promising drug target [37] – no PQS was found in the NS segment (which codes for the non-structural NS protein).

To evaluate the presence of the PQS in individual fragments we randomize five-times the RNA sequences of the Liaoming group (the group with highest PQS frequency) and as well in the Shandong group (the group with the lowest PQS frequency). A significant difference

Table 4 Summary of the in vitro G4 formation analyses by CD spectroscopy and ThT fluorescent assay in vitro. Sequences are shown in the 5′ to 3′ direction; all oligonucleotides are RNA. For G4 formation by CD, "Yes" indicates that a CD signature typical of a parallel G4 structure in the presence of K^+. The result of CD spectroscopy was considered positive in the case of a blue-shift of the positive ellipticity peak (from 270 to 264 nm) and a stronger signal in the presence of K^+ ions. Ratio between ThT fluorescence in the presence of oligonucleotide and background fluorescence of ThT alone is presented in the last column. The light-up effect (("Fold of ThT") refers to fold increase in Thioflavin T fluorescence emission when the candidate sequence is added: the higher this increase, the more likely is the structure to form a G4 motif.

Name	Sequence	G4Hunter score	G4 formation by CD	Fold of ThT
A0	GGGGAAGGAAGUGCUUGUAAUUUGGGG	1.407	Yes	7.23
A1	GGGAAAGGAAGUGCUUGUAAUUUGGGG	1.148	Yes	8.79
B0	UGGGGCAUUUUCCGAGGGUGACGGGA	1.154	Yes	7.90
C0	GAAGGGGGUAGGGACAAUGGUGAUGGAG	1.393	Yes	12.06
C1	AAAAGGAAUAGGAACAAUGGUAAUGGAG	0.571	No	6.23
D0	UGGGGCAAAUGGGGU	2.067	Yes	4.95
E0	GGCUUUGAGGGGGCCUGACGGG	1.318	No	4.28
E1	GGCUUUGAGGGGGCCUGACGGU	1.091	No	4.40
NK	UCUGACCGACCUAGAU	0.37	No	1.69

Fig. 5 Circular dichroism (CD) spectra of selected PQS in 1 mM sodium phosphate buffer (pH 7) (blue lines) or in 1 mM sodium phosphate (pH 7), 10 mM potassium phosphate (pH 7), 90 mM KCl buffer (orange lines); **a**. Oligonucleotide AO **b**. Negative control (NC); **c** Oligonucleotide CO; **d** Oligonucleotide C1

in PQS frequency was found between reference and randomized sequences for the NA segment of both groups (SM_05). On the other hand, the PQS frequency was not significantly different for other frogments, except for a depletion in the PA fragment in the Shandong group. These results are in agreement with recently proposed hypotheses that viruses causing acute infection are depleted in PQS (Bohálová N, Cantara A, Bartas M, Kaura P, Šťastný J, Pečinka P, Fojta M., Mergny J-L., Brázda V.: Analyses of viral genomes for G-quadruplex forming sequences reveal their correlation with the type of infection (submitted). A similar finding was published for SARS-CoV-2 [22]. On the other hand, the abundance of PQS in in the NA segment suggests its important evolutionarily conserved function.

Of note, none of the PQS identified here match a classical quadruplex consensus, in which four runs of three or more guanines are separated by 7 nucleotides or less, as predicted by Quadparser [38] with default parameters. As RNA G4 structures tend to be more stable than DNA, some of the motifs found here are still likely to form quadruplexes under physiological conditions, and this was experimentally confirmed using a combination of two biophysical methods. Given that all G4Hunter scores were relatively low, G4 formation was not a given, and needed the experimental confirmation. Our results show that the most conserved PQS in HA fragment, one with a conserved G run in NA fragment, as well as some

others are capable to form G-quadruplex structure in vitro, as shown by CD spectroscopy and by Fluorescence light-up measurements, while two sequence with low G4Hunter scores (< 0.6) did not form stable quadruplexes at room temperature. Interestingly, and as observed previously, the "grey zone" for which a sequence may well form a quadruplex or not seems to be centered around 1.1–1.2, and we have several sequences with relatively similar scores (between 1.09 and 1.31) which give different outcomes. G4Hunter is therefore not perfect – as all current prediction tools – and we are currently working on modifying parameters to improve accuracy. This may prove more difficult for RNA than for DNA, as we currently have access to far more experimental data on DNA than on RNA oligonucleotides.

In contrast to Quadparser, G4Hunter does not pick individual G-tracts to propose a core quadruplex with three loops. As demonstrated by a number of studies, the universe of G4-forming sequences is very diverse, and may involve bulges or snapback motifs, allowing individual, isolated guanines to participate in G-quadruplex formation. There are, of course, specific cases in which G participating to G-quartets can reasonably be assumed. For example, a bimolecular four-layer G4 motif can be predicted within the D motif for the GGGGCAAUGGGG region. In addition, for all motifs, one can imagine an intramolecular structure, provided

Brázda *et al. BMC Genomics*   (2021) 22:77

Page 9 of 11

that *i)* two-layered RNA G-quadruplexes are stable, and that *ii)* zero-nucleotide loops are allowed (such propeller loops have been found in a limited number of cases). In addition, one cannot exclude that isolated G also contribute to the core quartets. For this reason, in the absence of high resolution structures, it is rather premature to propose which G within these motifs are involved in G4 formation. These observations further illustrate that it is not possible to cover all G4-forming motifs with a single general consensus sequence.

These structures may offer opportunities for regulation and targeting by G4 ligands. Interestingly, several conserved PQS contain two GGGG runs, which may allow stable bimolecular G4 formation, as suggested for genome organization in other viruses including SARS-CoV-2 [39, 40].

Both strands, negative-sense genomic RNA and positive-sense mRNA, were analyzed for the presence and distribution of PQS as both RNAs are involved in lifecycle of the virus. Our result show that PQS are not evenly distributed but are mostly located on the RNA positive strand, thereby they may be involved in translation and splicing regulation. The genome of IAV is not stable and varies remarkably among strains [10, 41]. Comparison of PQS in various strains demonstrated that several PQS in the M segment and HA segment are highly conserved and therefore may be considered as suitable candidate targets with therapeutic potential. The HA segment (hemagglutinin) codes for a primary viral protein, which is recognized by the immune system and also is the primary target for vaccine design [42]. HA contains two subunits: HA1, which is responsible for receptor binding and HA2, which function is to support HA1 and mediates membrane fusion during viral entry [43]. Moreover, the conserved PQS in the HA sequence has the highest G4Hunter score among all found PQS. Another highly conserved PQS (just 1 variable nucleotide as shown in Table 3) was found at the 3′ end of an intron in negative-sense genomic RNA, near the splicing site of M2 protein. M1 is the only viral structural component which plays a major role in virus particle assembly [44]. M2 is a transmembrane ion channel protein which plays an important role in early stages of viral entry. Moreover, the M segment of 2009 H1N1 pandemic influenza virus was derived from the Eurasian avian-like swine lineage and was shown to affect neuraminidase activity and therefore might also have a potent effect on transmissibility [45].

Compared to strong depletion of PQS in the contemporarily outbreaking SARS-CoV-2 virus [22, 46], the genomes of IAV contain a highly conserved PQS, which could serve as a selective target. Our comprehensive analyses confirmed that several candidates adopt a quadruplex fold, arguing for the therapeutic potential of these PQS as targets for specific ligands.

Methods

Source of DNA sequences
The complete set of 616 sequences of 77 *G4*-EA-H1N1 genomes (each genome is divided into 8 segments) was downloaded on July 3, 2020 from the Genome database of the National center for Biotechnology Information (NCBI) [47]. NCBI accessions are shown in SM_01.

Process of analysis
All sequences belonging to 77 strains of *G4*-EA-H1N1 were analyzed with the G4Hunter Web tool [21], which is capable to read the NCBI identifier of the sequences uploaded in a .csv files. Default parameters for G4Hunter were set to "25" for window size and 1.2 or above for G4Hunter threshold score. G4Hunter score was grouped to the five intervals: 1.2–1.4, 1.4–1.6, 1.6–1.8, 1.8–2.0 and 2.0 and more, as previously performed for the Bacteria domain [19]. PQS frequencies were analyzed according to individual *G4*-EA-H1N1 strains, grouped according to regions of isolation and according to their eight genomic segments. All results including information about the size of genomic DNA sequence, number of PQS and statistical data are shown (SM_02A – grouped by region and SM_02B – grouped by segments).

Statistical evaluation
Statistical analysis of normality was made with the Shapiro-Wilk test. Since it was found that the data do not have a normal distribution, we used Kruskal-Wallis signed rank test to evaluate significant differences among strains and segments. Post-hoc multiple pairwise comparison by Dunn's test with Bonferroni correction of the significance level was applied with *p*-value cut-off 0.05. Data are available in SM_03.

The statistical significance analysis of the found PQS with respect to the same genome composition in a scrambled order was performed. The strains with the highest and lowest PQS frequency were selected. Reference sequences of individual segments of selected strains were five times randomized by the program Sequence Manipulation Suite – Shuffle DNA [48]. PQS frequencies were evaluated equally in reference and randomized sequences and plotted (SM_05).

Construction of LOGO sequences
NCBI sequences by list from SM_01 in FASTA format were downloaded and the dataset was uploaded to SnapGene (Align Multiple DNA Sequences) program. For every PQS we used the corresponding sequences from all *G4*-EA-H1N1 genomes and alignments by Clustal Omega tool [49] were generated. All found PQS were

searched in aligned sequences and WebLogo 3 [50] was used for generating LOGO sequences, all predicted PQS and LOGO sequences for PQS are available in SM_04.

CD spectroscopy

Synthetic oligonucleotides were purchased from Sigma-Aldrich and diluted in water. Oligonucleotides were heated at 85 °C for 3 min in 1 mM sodium phosphate buffer pH 7 or 1 mM sodium phosphate pH 7, 10 mM potassium phosphate pH 7 and 90 mM KCl buffer and slowly cooled down to room temperature. CD measurements were carried out with a Jasco 815 (Jasco International Co., Ltd.,Tokyo, Japan) dichrograph in 1 cm path-length microcells at 23 °C. A set of four scans with a data pitch of 0.5 nm and 200 nm/min scan speed was averaged for each sample. CD signal is expressed as the difference in molar absorption, $\Delta\varepsilon$, of the left- and right-handed circularly polarized light [31].

ThT fluorescent assay

Synthetic oligonucleotides were purchased from Sigma Aldrich and diluted in water. Oligonucleotides were diluted to 2 μM final concentration in 100 mM Tris-HCl pH 7.5 and 100 mM KCl buffer, heated at 85 °C for 3 min and slowly cooled down to room temperature. ThT was diluted in water to 1 μM final concentration and 100 mM KCl was added. Experiments were performed in a 384-well microplate from CORNING (Flat Bottom Black Polyester). Each condition was tested in triplicate. Measurements were performed at room temperature. Oligonucleotides and ThT were mixed at 0.5:1 M ratio to a final volume of 20 μl. Fluorescent emission was collected at 490 nm after excitation at 425 nm with a microplate reader (Spark, Tecan) [30].

Abbreviations
G4-EA-H1N1: Genotype 4 reassortant Eurasian avian-like H1N1 virus; PQS: Potential quadruplex-forming sequence; G4: Quadruplex; g-RNA: Viral RNA genome; IAV: Influenza A virus; HA: Hemagglutinin; NA: Neuraminidase; PB1: Polymerase basic protein 1; PB2: Polymerase basic protein 2; PA: Polymerase acidic protein; NP: Neuroprotein; M: The matrix protein; NS: The non-structural protein; CD: Circular dichroism; ThT: Thioflavin T

Acknowledgements
Not Applicable.

Authors' contributions
AC and NB performed analysis of dataset. OP collected and interpreted results of analysis and wrote the initial draft of the manuscript. NB and JC performed experimental methods. VB and MB revised the draft and wrote notes of the final content. JLM and MF designed the study and contributed to writing and critically edited the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the SYMBIT project (reg. no. CZ.02.1.01/0.0/0.0/15_003/0000477) financed by the ERDF.

Availability of data and materials
Data availability: all data are enclosed in the Supplementary materials (SM).

SM_01: NCBI accession numbers for all tested sequences.
SM_02: G4Hunter analyses results (A) strains (B) fragments.
SM_03: Statistical evaluation.
SM_04: PQS LOGO representation.
SM_05: Statistical analysis of randomized sequences (A) Shandong (B) Liaoning.
SM_06: Circular dichroism (CD) spectroscopy and the Thioflavin T (ThT) fluorescent assay results.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, 612 65 Brno, Czech Republic. [2]Faculty of Chemistry, Brno University of Technology, Purkyňova 118, 612 00 Brno, Czech Republic. [3]Department of Experimental Biology, Faculty of Science, Masaryk University, Kamenice 5, 62500 Brno, Czech Republic. [4]Department of Biology and Ecology/Institute of Environmental Technologies, Faculty of Science, University of Ostrava, 710 00 Ostrava, Czech Republic.

References
1. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza a epidemic. Nature. 2009;459(7250):1122–5.
2. Hoffmann M, Pöhlmann S. Cell entry of influenza a viruses: sweet talk between HA and CaV1.2. Cell Host Microbe. 2018;23(6):697–9.
3. Neumann G, Noda T, Kawaoka Y. Emergence and pandemic potential of swine-origin H1N1 influenza virus. Nature. 2009;459(7249):931–9.
4. Zhang R, Xu C, Duan Z. Novel antigenic shift in HA sequences of H1N1 viruses detected by big data analysis. Infect Genet Evol. 2017;51:138–42.
5. Mostafa A, Kanrai P, Ziebuhr J, Pleschka S. The PB1 segment of an influenza a virus H1N1 2009pdm isolate enhances the replication efficiency of specific influenza vaccine strains in cell culture and embryonated eggs. J Gen Virol. 2016;97(3):620–31.
6. Lee J-Y, Ouh I-O, Cho S-D, Cho I-S, Park CK, Song J-Y. Complete genome sequence of H1N1 swine influenza virus from pigs in the Republic of Korea in 2016. Baltrus DA, editor. Microbiol Resour Announc 2018;7(23):e01229–e01218, e01229-18.
7. Sullivan SJ, Jacobson RM, Dowdle WR, Poland GA. 2009 H1N1 Influenza. Mayo Clin Proc. 2010;85(1):64–76.
8. Zimmer SM, Burke DS. Historical perspective — emergence of influenza a (H1N1) viruses. N Engl J Med. 2009;361(3):279–85.
9. Shope RE. The incidence of neutralizing antibodies for swine influenza virus in the sera of human beings of different ages. J Exp Med. 1936;63(5):669–84.
10. Sun H, Xiao Y, Liu J, Wang D, Li F, Wang C, et al. Prevalent Eurasian avian-like H1N1 swine influenza virus with 2009 pandemic viral genes facilitating human infection. Proc Natl Acad Sci U S A. 2020;29:201921186.
11. Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S. Quadruplex DNA: sequence, topology and structure. Nucleic Acids Res. 2006;34(19):5402–15.
12. Malgowska M, Czajczynska K, Gudanis D, Tworak A, Gdaniec Z. Overview of the RNA G-quadruplex structures. Acta Biochim Pol. 2016;63(4):609–21.
13. Varshney D, Spiegel J, Zyner K, Tannahill D, Balasubramanian S. The regulation and functions of DNA and RNA G-quadruplexes. Nat Rev Mol Cell Biol. 2020 Aug;21(8):459–74.
14. Ruggiero E, Richter SN. Viral G-quadruplexes: New frontiers in virus pathogenesis and antiviral therapy. In: Annual Reports in Medicinal Chemistry. Elsevier; 2020 [cited 2020 Jul 29]. p. S0065774320300142.
15. Lavezzo E, Berselli M, Frasson I, Perrone R, Palù G, Brazzale AR, et al. G-quadruplex forming sequences in the genome of all known human viruses: A comprehensive guide. PLoS Comput Biol. 2018 13; 14(12).

16. Lombardi EP, Londoño-Vallejo A. A guide to computational methods for G-quadruplex prediction. Nucleic Acids Res. 2020;48(3):1603.

17. Bedrat A, Lacroix L, Mergny J-L. Re-evaluation of G-quadruplex propensity with G4Hunter. Nucleic Acids Res. 2016;44(4):1746–59.

18. Gazanion E, Lacroix L, Alberti P, Gurung P, Wein S, Cheng M, et al. Genome wide distribution of G-quadruplexes and their impact on gene expression in malaria parasites. Di Antonio M, editor. PLoS Genet. 2020;16(7):e1008917.

19. Bartas M, Čutová M, Brázda V, Kaura P, Šťastný J, Kolomazník J, et al. The presence and localization of G-Quadruplex forming sequences in the domain of Bacteria. Molecules. 2019;24(9):1711.

20. Čutová M, Manta J, Porubiaková O, Kaura P, Šťastný J, Jagelská EB, et al. Divergent distributions of inverted repeats and G-quadruplex forming sequences in Saccharomyces cerevisiae. Genomics. 2019: S0888754319305269.

21. Brázda V, Kolomazník J, Lýsek J, Bartas M, Fojta M, Šťastný J, et al. G4Hunter web application: a web server for G-quadruplex prediction. Hancock J, editor. Bioinformatics. 2019;35(18):3493–5.

22. Bartas M, Brázda V, Bohálová N, Cantara A, Volná A, Stachurová T, et al. In-depth Bioinformatic analyses of Nidovirales including human SARS-CoV-2, SARS-CoV, MERS-CoV Viruses Suggest Important Roles of Non-canonical Nucleic Acid Structures in Their Lifecycles. Front Microbiol. 2020;11:1583.

23. Di Antonio M, Ponjavic A, Radzevičius A, Ranasinghe RT, Catalano M, Zhang X, et al. Single-molecule visualization of DNA G-quadruplex formation in live cells. Nat Chem [Internet]. 2020 20.

24. Prorok P, Artufel M, Aze A, Coulombe P, Peiffer I, Lacroix L, et al. Involvement of G-quadruplex regions in mammalian replication origin activity. Nat Commun. 2019;10(1):3274.

25. Artusi S, Perrone R, Lago S, Raffa P, Di Iorio E, Palù G, et al. Visualization of DNA G-quadruplexes in herpes simplex virus 1-infected cells. Nucleic Acids Res. 2016;44(21):10343–53.

26. Callegaro S, Perrone R, Scalabrin M, Doria F, Palù G, Richter SN. A core extended naphtalene diimide G-quadruplex ligand potently inhibits herpes simplex virus 1 replication. Sci Rep. 2017;7(1):2341.

27. Métifiot M, Amrane S, Litvak S, Andreola M-L. G-quadruplexes in viruses: function and potential therapeutic applications. Nucleic Acids Res. 2014;42(20):12352–66.

28. González VM, Martín ME, Fernández G, García-Sacristán A. Use of Aptamers as Diagnostics Tools and Antiviral Agents for Human Viruses. Pharmaceuticals. 2016;9(4):78.

29. Majerciak V, Zheng Z-M. Detection of Viral RNA Splicing in Diagnostic Virology. In: Tang Y-W, Stratton CW, editors. Advanced Techniques in Diagnostic Microbiology. Cham: Springer International Publishing; 2018. p. 345–402.

30. Renaud de la Faverie A, Guédin A, Bedrat A, Yatsunyk LA, Mergny J-L. Thioflavin T as a fluorescence light-up probe for G4 formation. Nucleic Acids Res. 2014;42(8):e65.

31. Vondrušková J, Kypr J, Kejnovská I, Fialová M, Vorlíčková M. Guanine quadruplex formation by RNA/DNA hybrid analogs of Oxytricha telomere G4T4G4 fragment. Biopolymers. 2008;89(10):797–806.

32. Brázda V, Laister RC, Jagelská EB, Arrowsmith C. Cruciform structures are a common DNA feature important for regulating biological processes. BMC Mol Biol. 2011;12(1):33.

33. Ruggiero E, Richter SN. G-quadruplexes and G-quadruplex ligands: targets and tools in antiviral therapy. Nucleic Acids Res. 2018;46(7):3270–83.

34. Shen L-W, Qian M-Q, Yu K, Narva S, Yu F, Wu Y-L, et al. Inhibition of influenza a virus propagation by benzoselenoxanthenes stabilizing TMPRSS2 gene G-quadruplex and hence down-regulating TMPRSS2 expression. Sci Rep. 2020;10(1):7635.

35. Eisfeld AJ, Neumann G, Kawaoka Y. At the Centre: influenza a virus ribonucleoproteins. Nat Rev Microbiol. 2015;13(1):28–41.

36. Compans RW, Content J, Duesberg PH. Structure of the ribonucleoprotein of influenza virus. J Virol. 1972;10(4):795–800.

37. Hu Y, Sneyd H, Dekant R, Wang J. Influenza A Virus Nucleoprotein: A Highly Conserved Multi-Functional Viral Protein as a Hot Antiviral Drug Target. Curr Top Med Chem. 2017;17(20):2271–85.

38. Kikin O, D'Antonio L, Bagga PS. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Res. 2006; 34(Web Server):W676–W682.

39. Dolinnaya NG, Ogloblina AM, Yakubovskaya MG. Structure, properties, and biological relevance of the DNA and RNA G-quadruplexes: overview 50 years after their discovery. Biochem Moscow. 2016;81(13):1602–49.

40. Hognon C, Miclot T, García-Iriepa C, Francés-Monerris A, Grandemange S, Terenzi A, et al. Role of RNA guanine Quadruplexes in Favoring the dimerization of SARS unique domain in coronaviruses. J Phys Chem Lett 2020;11(14):5661–5667.

41. Speranskaia AS, Mel'nikova NV, Belenkin MS, Dmitriev AA, Oparina NI, Kudriavtseva AV. Genetic diversity and evolution of the influenza C virus. Genetika. 2012 Jul;48(7):797–805.

42. Xu R, Ekiert DC, Krause JC, Hai R, Crowe JE, Wilson IA. Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. Science. 2010;328(5976):357–60.

43. Kuenstling TE, Sambol AR, Hinrichs SH, Larson MA. Oligomerization of bacterially expressed H1N1 recombinant hemagglutinin contributes to protection against viral challenge. Sci Rep. 2018;8(1):11856.

44. Gómez-Puertas P, Albo C, Pérez-Pastrana E, Vivo A, Portela A. Influenza virus matrix protein is the major driving force in virus budding. J Virol. 2000; 74(24):11538–47.

45. Campbell PJ, Danzy S, Kyriakis CS, Deymier MJ, Lowen AC, Steel J. The M segment of the 2009 pandemic influenza virus confers increased neuraminidase activity, filamentous morphology, and efficient contact transmissibility to a/Puerto Rico/8/1934-based Reassortant viruses. J Virol. 2014;88(7):3802–14.

46. Ji D, Juhas M, Tsang CM, Kwok CK, Li Y, Zhang Y. Discovery of G-quadruplex-forming sequences in SARS-CoV-2. Brief Bioinform. 2020: bbaa114.

47. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, et al. Database resources of the National Center for biotechnology information. Nucleic Acids Res. 2019;47(D1):D23–8.

48. Stothard P. The sequence manipulation suite: JavaScript programs for Analyzing and formatting protein and DNA sequences. BioTechniques. 2000; 28(6):1102–4.

49. Sievers F, Higgins DG. Clustal omega. Curr Protoc Bioinformatics. 2014;48:3. 13.1–16.

50. Crooks GE. WebLogo: A Sequence Logo Generator. Genome Res. 2004;14(6): 1188–90.

Publisher's Note