PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

G-quadruplex forming sequences in the genome of all known human viruses: A comprehensive guide

Enrico Lavezzo[1☯], **Michele Berselli**[1☯], **Ilaria Frasson**[1☯], **Rosalba Perrone**[1], **Giorgio Palù**[1], **Alessandra R. Brazzale**[2]*, **Sara N. Richter**[1]*, **Stefano Toppo**[1]*

1 Department of Molecular Medicine, University of Padova, Padova, Italy, **2** Department of Statistical Sciences, University of Padova, Padova, Italy

☯ These authors contributed equally to this work.
* alessandra.brazzale@unipd.it (ARB); sara.richter@unipd.it (SNR); stefano.toppo@unipd.it (ST)

Citation: Lavezzo E, Berselli M, Frasson I, Perrone R, Palù G, Brazzale AR, et al. (2018) G-quadruplex forming sequences in the genome of all known human viruses: A comprehensive guide. PLoS Comput Biol 14(12): e1006675. https://doi.org/10.1371/journal.pcbi.1006675

Editor: Matej Lexa, Masarykova Univerzita, CZECH REPUBLIC

Received: August 7, 2018

Accepted: November 27, 2018

Published: December 13, 2018

Data Availability Statement: The data underlying the results presented in the study are available from http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=g4virus.

Funding: This work was supported by the Bill and Melinda Gates Foundation [GCE grant numbers OPP1035881, OPP1097238 to SNR]; European Research Council [ERC Consolidator grant 615879 to SNR]; Italian Ministry of University and Research [FIRB-Ideas RBID082ATK to SNR]; and University of Padova [PRAT CPDA138081/13 to

Abstract

G-quadruplexes are non-canonical nucleic-acid structures that control transcription, replication, and recombination in organisms. G-quadruplexes are present in eukaryotes, prokaryotes, and viruses. In the latter, mounting evidence indicates their key biological activity. Since data on viruses are scattered, we here present a comprehensive analysis of potential quadruplex-forming sequences (PQS) in the genome of all known viruses that can infect humans. We show that occurrence and location of PQSs are features characteristic of each virus class and family. Our statistical analysis proves that their presence within the viral genome is orderly arranged, as indicated by the possibility to correctly assign up to two-thirds of viruses to their exact class based on the PQS classification. For each virus we provide: i) the list of all PQS present in the genome (positive and negative strands), ii) their position in the viral genome, iii) the degree of conservation among strains of each PQS in its genome context, iv) the statistical significance of PQS abundance. This information is accessible from a database to allow the easy navigation of the results: http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=g4virus. The availability of these data will greatly expedite research on G-quadruplex in viruses, with the possibility to accelerate finding therapeutic opportunities to numerous and some fearsome human diseases.

Author summary

G-quadruplexes are nucleic acid non-canonical structures that have been implicated in the regulation of different biological processes of many organisms. Their presence has been demonstrated also in several viral pathogens, providing new insights into viruses' biology and potentially serving as drug targets. Although experimental validation is needed to confirm the actual folding of G-quadruplexes, they can be inferred *in silico* directly from the nucleotide sequence. Several computational methods exist for this purpose, but they are all limited to the analysis of independent sequences. Since viral genomes can be highly variable, G-quadruplexes with important functional roles are expected to be

ST]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

conserved among strains and isolates belonging to the same viral species. Here we aimed at characterizing the potential quadruplex-forming sequences (PQS) content in the genome of viral human pathogens and assess their degree of conservation in each viral species. We demonstrate that many viruses possess more PQSs than expected from their nucleotide composition and some of them are highly conserved within single viral species, claiming some biological roles. We provide a website where the results of our analyses are displayed for each virus with interactive graphics. This work is intended as a resource that can guide scientists in the choice of the most promising candidates for functional characterization.

Introduction

G-quadruplexes (G4s) are nucleic-acid secondary structures that may form in single-stranded DNA and RNA G-rich sequences under physiological conditions [1]. Four Gs bind via Hoogsteen-type base-pairing to yield G-quartets: stacking of at least two G-quartets leads to G4 formation, through π-π interactions between aromatic systems of G-quartets. K^+ cations in the central cavity relieve repulsion among oxygen atoms and specifically support G4 formation and stability [2]. In the human genome, potential quadruplex-forming sequences (PQS) are clustered at definite genomic regions, such as telomeres, oncogene promoters, immunoglobulin switch regions, DNA replication origins and recombination sites [3]. In RNA, G4s and PQSs were mapped in mRNAs and in non-coding RNAs (ncRNAs) [4], such as long non-coding RNAs (lncRNAs) [5] and precursor microRNAs (pre-miRNAs) [6] indicating the potential of RNA G4s to regulate both pre- and post-transcriptional gene expression [7, 8].

Viruses are intracellular parasites that replicate by exploiting the cell replication and protein synthesis machineries. Viruses that infect humans are very diverse and, according to the Baltimore classification, they can be divided in seven groups based on the type of their genome and mechanism of genome replication: DNA viruses with 1) double-stranded (ds) and 2) single-stranded (ss) genome; RNA viruses with 3) ds genome, or ss genome with 4) positive (ssRNA (+)) or 5) negative (ssRNA (-)) polarity; 6) RNA or 7) DNA viruses with reverse transcription (RT) ability, whose genome is converted from RNA to DNA during the virus replication cycle (Table 1). Each of these classes possesses a peculiar replication cycle [9].

The presence of G4s in viruses and their involvement in virus key steps is increasingly evident in most of the Baltimore groups [10, 11]. In the dsDNA group, G4s were described in both *Herpesviridae* and *Papillomaviridae* families [12–20]. In ssDNA viruses, the presence of

Table 1. Virus families.

Genome nature	DNA		RNA			DNA and RNA	
Group	1	2	3	4	5	6	7
Genome type	dsDNA	ssDNA	dsRNA	ssRNA (+)	ssRNA (-)	ssRNA (RT)	dsDNA (RT)
Virus family	Herpes	Anello	Reo	Corona	Rhabdo	Retro	Hepadna
	Adeno	Parvo		Astro	Filo		
	Papilloma			Calici	Paramyxo		
	Polyoma			Flavi	Arena		
	Pox			Picorna	Bunya		
				Toga	Orthomyxo		

Virus families divided according to their genome and mechanism of replication. The suffix word "viridae" for each virus family has been omitted.

https://doi.org/10.1371/journal.pcbi.1006675.t001

G4s was reported in the adeno-associated virus genome [21]. RNA G4s were described in the genomes of both ssRNA (+) (i.e. Zika, hepatitis C virus (HCV) [22, 23], and the severe acute respiratory syndrome (SARS) coronavirus [24, 25]) and ssRNA (-) viruses (i.e. Ebola virus [26]). A G4 was also detected in hepatitis B virus (HBV) genome, the only member of dsDNA viruses with RT activity [27]. Finally, functionally significant G4s were identified both in the RNA and DNA proviral genome of the human immunodeficiency virus (HIV), a retrovirus belonging to group 6 (Table 1) [28–35], and [33, 34]in the LTR region of lentiviruses in general (ssRNA RT) [36].

Given this amount of scattered data, we here aimed at analyzing the presence of PQSs in the genome of all known viruses that can cause infections in humans. The analysis is performed at two distinct levels, globally for each viral genome and individually for each detected PQS. We asked the following: is the number of PQSs found in a viral genome simply due to chance, hence trivially reflecting genomic G/C content? And how much is each PQS conserved among the strains belonging to a viral species? To address these questions, we collected the whole viral genomes deposited in databanks, scanned them to detect all PQSs, and performed different statistical evaluations following the data analysis workflow shown in Fig 1. The detailed information on PQSs present in each human virus is available in an easily accessible web site with interactive graphics and genome browser visualization tools (http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=g4virus).

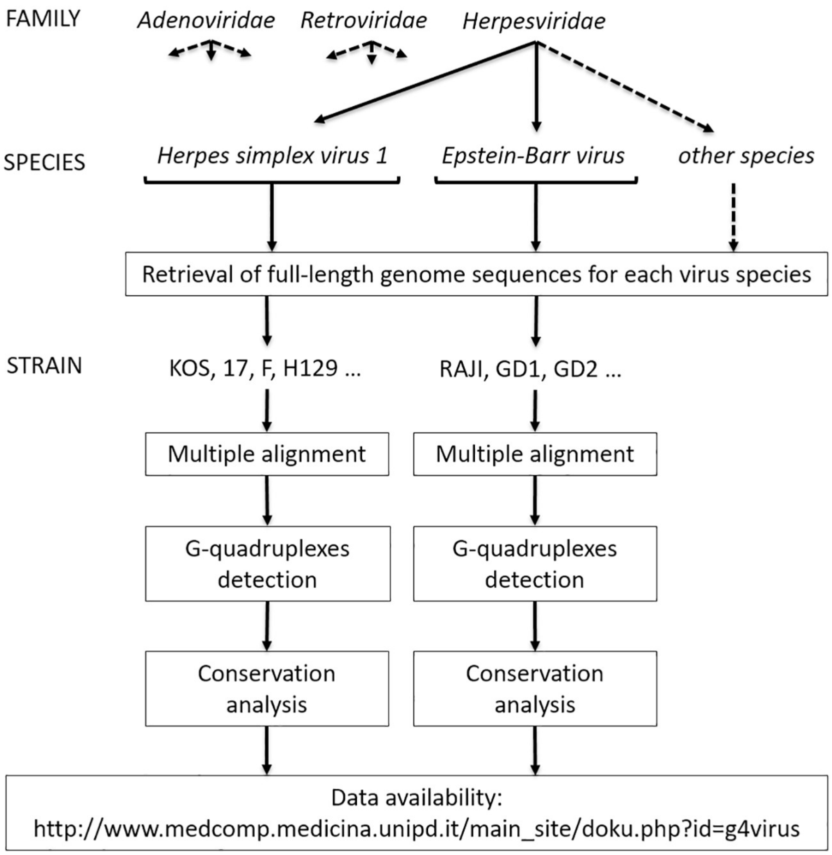

Fig 1. Example of virus classification in families, species and strains with data analysis flowchart. The conservation analysis was performed among the strains within each virus species.

https://doi.org/10.1371/journal.pcbi.1006675.g001

Results

Detection of G4 patterns in all known human viruses

All known viruses that cause infections in humans, according to the Viral Zone ExPASy web site (http://viralzone.expasy.org/all_by_species/678.html), were grouped in 7 classes according to Baltimore classification, which takes into account the viral genome nature: dsDNA, ssDNA, dsRNA, ssRNA(+), ssRNA(-), ssRNA(RT) and dsDNA(RT). Different replication strategies and structural similarities allow to further divide viruses in families (Table 1). The complete list of reference sequences for each virus included in the analyses is reported in S1 Table.

PQSs in viral genomes were searched by looking for the following patterns: [G(2)N(1–7)](3)G(2), [G(3)N(1–12)](3)G(3) and [G(4)N(1–12)](3)G(4), where both island and loop lengths were chosen to provide a comprehensive detection. We decided to expand the search to PQSs with very short islands and quite extended loops for the following reasons: first, the folding of PQS with GG-islands has been previously demonstrated in viruses [32]; second, since many viruses possess a RNA genome, and considering that RNA G4s are more stable than their DNA counterparts [37], PQSs with only two tetrads have a reasonable chance to fold in viral RNA genomes or in their intermediates. Finally, while long loops are known to destabilize G4 structures, their presence is anyway compatible with the folding of stable G4s at physiological temperature [38]. PQSs with bulged islands [39] and intermolecular G4s are not considered in the present study.

PQSs were searched in the positive and negative strand of each virus genome sequence, since both filaments are present and important in different stages of the viral replicative cycle of all virus classes. As the length of virus genomes greatly varies, i.e. from 235,646 nucleotides (nts) of the human cytomegalovirus (HCMV) to 1,682 nts of hepatitis delta virus (HDV), we reported the number of PQS independently of the genome length by normalizing their number per 1,000 nts (Fig 2). The PQS distribution for both the positive and negative strands is shown as a box plot for each Baltimore virus class, whereas the PQS count for each virus within each class is shown as a dot besides the box plot (Fig 2). The negative strand of retroviruses (ssRNA (RT) viruses), ssDNA viruses and both strands of dsDNA viruses showed the largest presence of PQSs made of GG-, GGG- and GGGG-islands (box plots, Fig 2). Both strands of genomes of single virus families belonging to these groups and to ssRNA (+) and ssRNA (-) were enriched in PQSs of all G-islands types (dot plots, Fig 2). Conversely, dsRNA and dsDNA (RT) viruses notably lacked the presence of PQSs.

Then, we evaluated the conservation of PQSs among different strains of each viral species, hypothesizing that the presence of a conserved PQS within a less conserved genome environment could be an indication of a G4 with a biological function [40]. To allow for the evaluation of PQS conservation in the local context of viral genomes, we computed the “G4 scaffold conservation index” (G4_SCI) for each PQS in each virus species. This value measures the degree of conservation of G-islands that are necessary and sufficient to form a PQS: the higher the score, the higher the conservation of the PQS. An example of the results from such analysis is reported in Fig 3 for the lymphocytic choriomeningitis virus (segment S): all PQSs detected in the virus are plotted as vertical bars, the height and position of which represent the G4_SCI on the y-axis and the genome coordinates on the x-axis, respectively. In addition, the local sequence conservation (LSC) of the viral genome, calculated with a sliding window approach on all available viral sequences, is reported alongside as a red broken line. This visualization method allows the prompt identification of the presence, position, and conservation of G-islands within PQSs, together with the overall local conservation of the genomic context. Moreover, the degree of conservation of the connecting regions (loops) with respect to G-islands (the *loop_conservation* value) was calculated as the difference between G4_SCI and

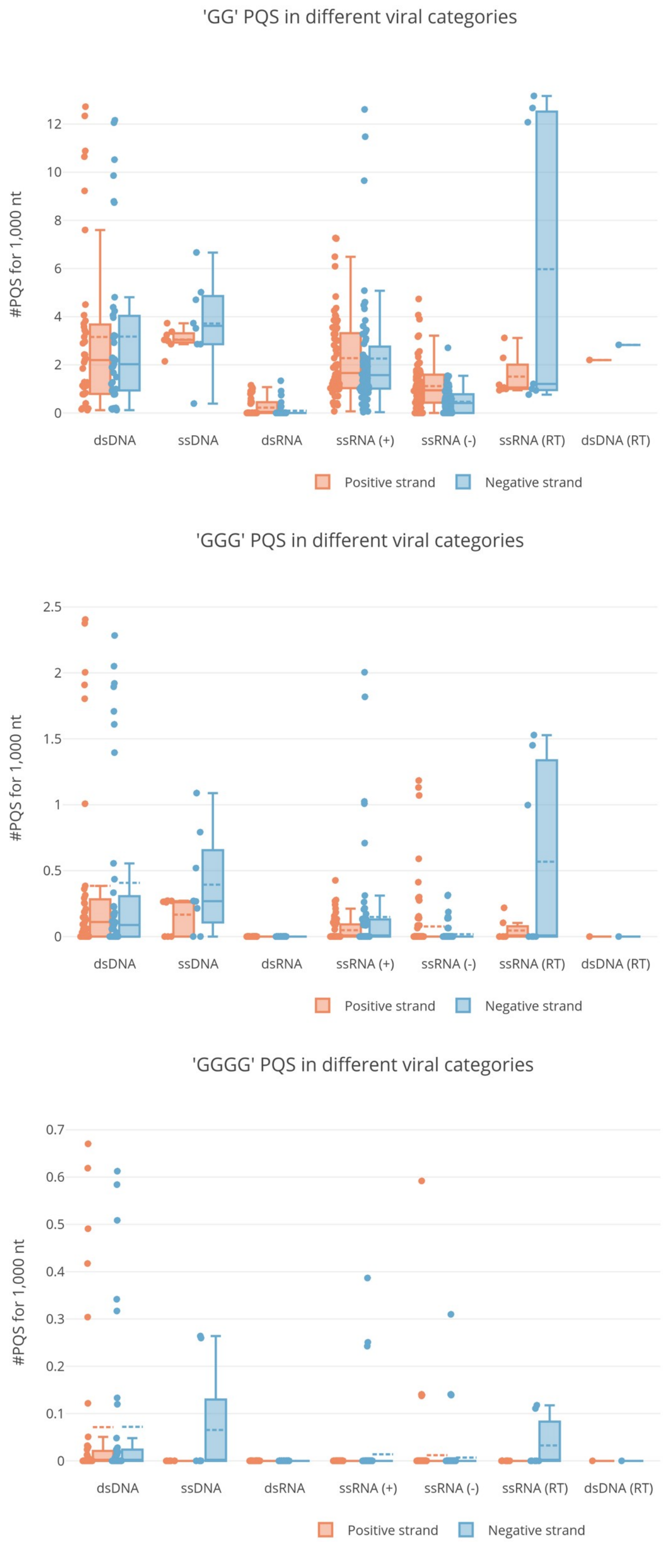

Fig 2. Box and whisker plots of PQSs in different virus classes. Each panel refers to the indicated type of G-island (GG, GGG, GGGG). The abundance of PQSs per 1 kb of viral genome is reported in the y-axis (for each viral species, the median value among all available strains is used) and the different virus categories in the x-axis. Boxplots are delimited by the first and third quartile and the straight and dotted lines drawn inside are the median and mean values, respectively, of the PQS distribution. The single observations are reported as dots close to the box plot. Whiskers

delimit all the points that fall above/below the third/first quartile plus/minus 1.5 times the interquartile range (IQR). Orange and blue box plots refer to positive and negative strand respectively.

https://doi.org/10.1371/journal.pcbi.1006675.g002

LSC. Positive and negative *loop_conservation* scores indicate, respectively, lower and higher conservation of connecting regions compared to the conservation of G-islands. Values close to zero mean that both G-islands and connecting loops show the same level of sequence conservation. In Fig 3, three PQSs formed by highly conserved GG-islands are shown for the S segment of lymphocytic choriomeningitis virus, present in genomic regions both well and less well conserved (Fig 3 at positions 1,790 in the positive strand, 1,760 and 2,680 in the negative strand). This kind of analysis is available for all PQSs of all human virus species at http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=g4virus (*loop_conservation* values are included in tarballs downloadable for each viral class of the Baltimore classification, whereas each virus species has a dedicated page displaying all graphical representations).

To assess the results, we retrieved from the literature all the available experimentally validated G4s detected in human viruses. All patterns were confirmed also by our analysis and the complete list is reported in S2 Table, together with the genomic coordinates of the predicted PQSs.

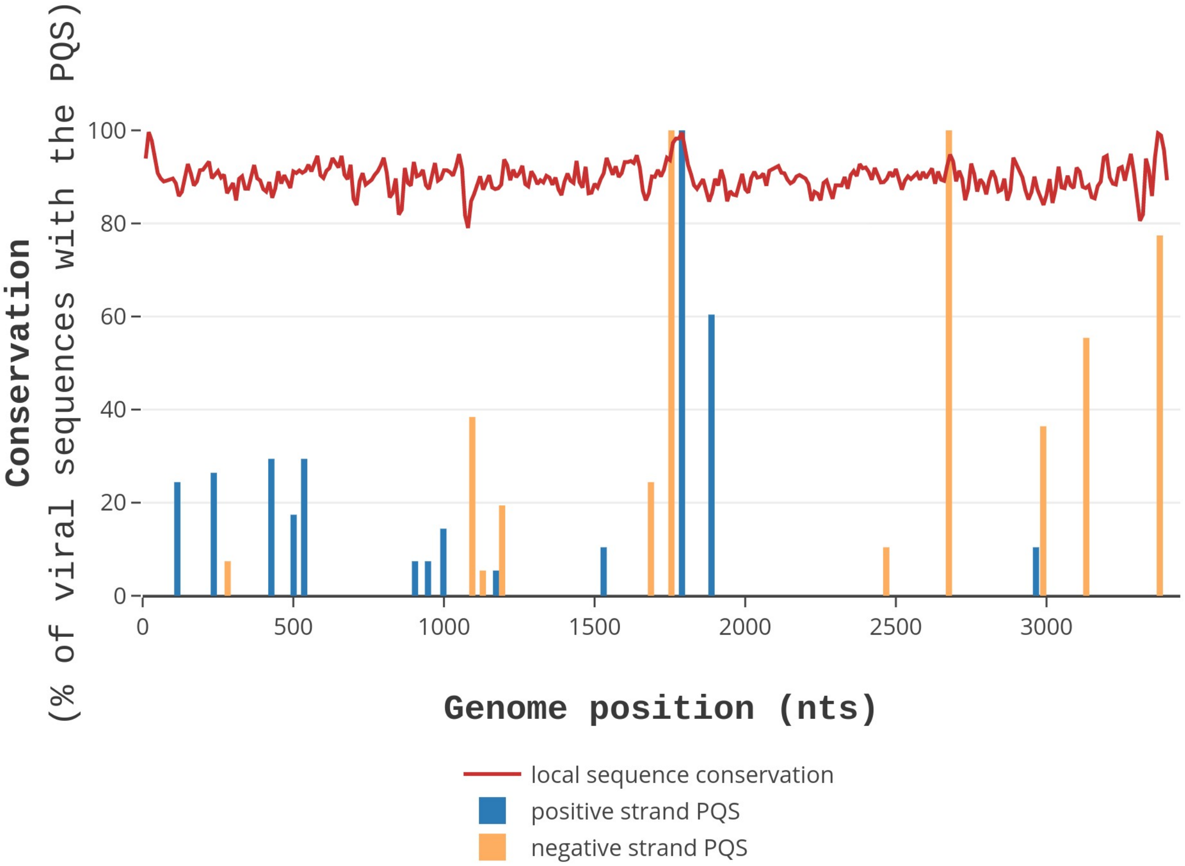

Fig 3. Conservation of PQSs and viral genomes. PQSs formed by GG-islands in the S segment of lymphocytic choriomeningitis virus are shown. PQSs found in the positive and negative strands are indicated as blue and orange vertical bars, respectively, while the height of bars represents the G4_SCI (the conservation of G-islands). Local sequence conservation (LSC) of viral genomes is shown as a red broken line. The x-axis indicates genome position, the y-axis the conservation %.

https://doi.org/10.1371/journal.pcbi.1006675.g003

Statistical evidence of the presence of PQSs in the human virus genomes

G4 formation may be largely affected by G/C content, which greatly varies in viral genomes (from 76% of Cercopithecine 2 herpes virus to 27% of Yaba like disease virus). Moreover, it has been shown that some di- and trinucleotides are over- or under-represented in certain viruses [41, 42] and, in the context of PQSs, this means that their abundance could be biased by unexpected frequencies of guanine homopolymers. G-island frequencies higher or lower than expected would lead to a potential over- or under-representation of PQSs, respectively.

To check whether the presence of PQSs was statistically relevant or whether it occurred by pure chance, we compared the results obtained from real viral genomes with those obtained by two different simulation strategies. The first one (single nucleotide assembling) assumes that the occurrence of each DNA base in the genome is independent [43]; the second (G-island reshuffling) considers that short sequences of a given length (k-mer) could be over- or under-represented in certain viral genomes [41, 42]. In the former case, sequences were generated with the same composition of nucleotides but different order with respect to references; in the latter, sequences were produced by reshuffling the positions of G-islands while keeping constant their number.

For each virus and simulation strategy, we produced 10,000 random sequences, which were screened with our PQSs detection pipeline. Real and simulated data were compared by computing a P-value, defined as twice the smaller proportion of simulated sequences that exhibit, respectively, a higher and lower count of PQSs as compared to the median value of all the available complete genome sequences for a certain virus. Hence, a P-value close to 1 means that the median PQS content in real viral sequences is not significant if compared to a random distribution; conversely, a P-value close to 0 means that PQS content is highly significant. This interpretation holds independently of the length of the genome and/or of the prevalence of either G/C bases or G-islands, as we compare the number of PQSs in a viral genome with the one we would expect in a simulated genome of the same length and of either the same base or G-island composition. To account for possible high discreteness of the data, a less conservative version of the P-value, called the mid-P value [44], was used. Segment diagrams of the mid-P values of the Baltimore grouped viruses are reported in S1 and S2 Figs [45]. The number of viruses whose median PQS count is significant at the 10% level is listed in Table 2 (virus names in S3 Table) with the indication of whether this median count is either higher or lower than the PQS count in simulated sequences.

Our data show that most members of the dsDNA, ssDNA, and ssRNA (RT) present a highly significant content of PQSs formed by GG-, GGG- and/or GGGG-islands in one or both strands. ssRNA (-) and ssRNA (+) classes are heterogeneous since some viruses are highly significant in any PQS category (from GG- to GGGG-islands), while others are not (see below). The presence of PQSs in members of the dsRNA group is notably less significant. Interestingly, few viruses display a smaller amount of PQSs than expected: both *Sagiyama virus* and *Human coronavirus HKU1* are depleted of PQSs belonging to GG-islands category in the positive genome strand when compared with both simulation strategies based on single nucleotide assembling and GG-island reshuffling. In addition, *Human parainfluenza virus 2* is poor of PQSs made of GG-island in the positive genome strand but is enriched in both GG- and GGG-type PQSs in the negative strand.

Overall, if we consider the viruses that contain at least one PQS in either the real or the simulated genomes, we observe that the increase in G-islands' length corresponds to a decrease in the absolute number of viruses containing PQSs, but it also corresponds to a dramatic increase in the fraction of them that is statistically significant.

Table 2. Relative abundance of viruses having a PQS content significantly different between real and simulated viral genomes.

	Number of viruses significantly different vs. randomization at single nucleotide level		Number of viruses significantly different vs. randomization at the G-island level		Number of viruses significantly different vs. both randomization	
G-island pattern	**PQS more abundant in real sequences**	**PQS more abundant in simulated sequences**	**PQS more abundant in real sequences**	**PQS more abundant in simulated sequences**	**PQS more abundant in real sequences**	**PQS more abundant in simulated sequences**
GG (positive strand)	83/218 (38.1%)	9/218 (4.1%)	52/217 (24.0%)	4/217 (1.8%)	49/217 (22.6%)	3/217 (1.4%)
GG (negative strand)	83/187 (44.4%)	2/187 (1.1%)	67/187 (35.8%)	1/187 (0.5%)	59/186 (31.7%)	0 (0%)
GGG (positive strand)	41/78 (52.6%)	3/78 (3.8%)	40/75 (53.3%)	0 (0%)	34/74 (45.9%)	0 (0%)
GGG (negative strand)	32/68 (47.1%)	3/68 (4.4%)	32/69 (46.4%)	0 (0%)	28/68 (41.2%)	0 (0%)
GGGG (positive strand)	17/19 (89.5%)	0 (0%)	17/17 (100%)	0 (0%)	17/17 (100%)	0 (0%)
GGGG (negative strand)	23/28 (82.1%)	0 (0%)	23/25 (92.0%)	0 (0%)	23/25 (92.0%)	0 (0%)

The number of viruses where the amount of PQSs is significantly different at 10% level between real and simulated sequences is reported (with percentages in brackets). Values and percentages were calculated considering only viruses containing at least one PQS either in real or simulated sequences (this explains differences in denominators). The table reports significant values for either one of the two simulations (randomization of viral genomes at single nucleotide or at G-island levels) or both.

https://doi.org/10.1371/journal.pcbi.1006675.t002

By looking at the family level of viral classification, which is far more homogeneous than the Baltimore groups, some virus families emerge as prominently enriched in PQSs. Among them, *Herpesviridae* is not only the one with the highest PQS content, but most of its members display significantly more PQSs than expected in both genome strands and in all considered G-island lengths. Notably, some of the viruses belonging to *Herpesviridae* and showing the highest G/C content are statistically enriched in PQSs. This suggests that simply having a high G/C content is not a sufficient condition to justify the presence of such a high number of PQSs. Other viral families that are consistently enriched in PQSs are *Adenoviridae* and *Papillomaviridae*, especially in GG- (both strands) and GGG-island (positive strand) types. *Poxviridae* and *Parvoviridae* show an enrichment of GG-type PQSs in both genome strands, whereas the same pattern is enriched in the positive strand of all *Anelloviridae* members and in the negative strand of most *Paramyxoviridae* and *Retroviridae* viruses. All other families are generally not enriched in PQSs in any of the evaluated categories, with only a few exceptions that are listed in the following: L segments of *Lassa virus* and *Lymphocytic choriomeningitis virus* (*Arenaviridae*), *Wu* and *Merkel cell polyomaviruses* (*Polyomaviridae*), *Salivirus* (*Picornaviridae*), M and S segments of respectively *Crimean-Congo hemorrhagic fever virus* and *Rift Valley fever virus* (*Bunyaviridae*).

By comparing the results obtained independently from the two simulation strategies it is possible to draw additional conclusions. First, in most cases the results are concordant, meaning that both simulations show similar trends in the statistical significance. Nonetheless, the overall number of viruses whose PQS content is significantly different with respect to simulated data is higher when real viral genomes are compared to those generated by single nucleotide assembling. This difference indicates that viral genome k-mer composition is indeed affecting the probability of randomly finding PQSs, at least in a proportion of viruses as shown in Fig 4: in the heatmaps, viruses that are significant in only one of the two simulations are

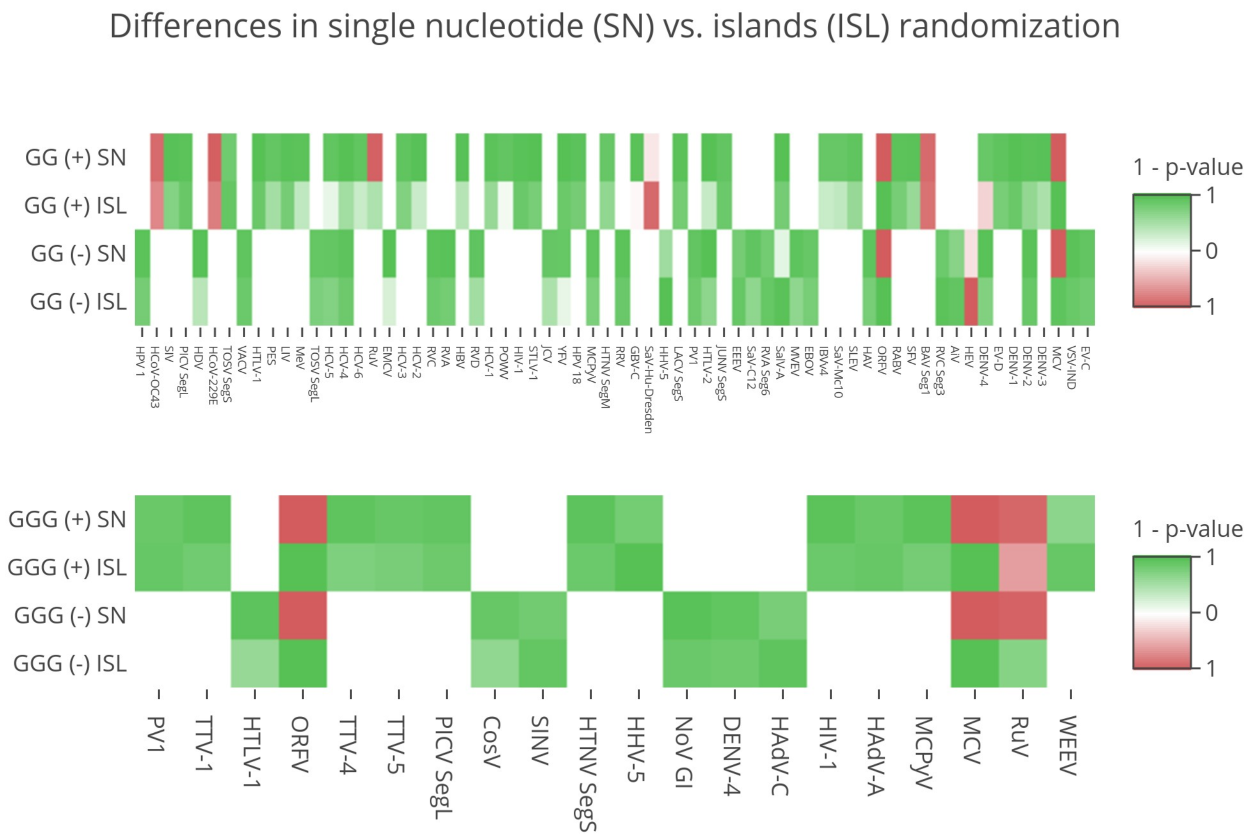

Fig 4. Different results from single nucleotide (SN) and islands (ISL) reshuffling strategies. Heatmaps show all the viruses which are significant in only one of the two simulations or that obtain discordant results. Green and red boxes indicate that PQSs are more abundant in real and simulated genomes, respectively, with color intensity proportional to p-value size.

https://doi.org/10.1371/journal.pcbi.1006675.g004

reported for GG- and GGG-island patterns, whereas no such cases were found for GGGG-type PQSs.

Finally, some remarkable exceptions exist where both simulations return a significant p-value, but with an opposite meaning. This is the case of two members of the *Poxviridae* family, namely *Molluscum contagiosum virus* and *Orf virus*, which are enriched in GG- and GGG-type PQSs in both strands of their genomes if compared with the islands reshuffling simulation but show the opposite behavior when compared with the single nucleotide assembling (they are also reported in Fig 4). While the full meaning of this observation is not clear to us, it seems that these viruses possess far less PQSs than they could have, but at the same time they are able to cluster their relatively few G-islands in more PQSs than expected.

Human virus PQSs position and overlap with genomic features

To check the prevalent positions of PQSs in virus genomes, we compared the coordinates of predicted PQSs with the available information regarding viral genome features. Genome coordinates were extracted for coding sequences (CDS), repeat regions (RR), 5'- and 3'-untranslated (UTR), and promoter regions. While CDS and RR are explicitly defined in RefSeq and GenBank databases, the annotation of UTRs and promoters is more inconsistent, being defined only for some viral species. For this reason, the annotations of genes and CDSs were exploited to indirectly extract the coordinates of 5'–and 3'–regulatory regions (see Materials and methods for details). To determine the localization of PQSs in viral genomes, the overlap extent between PQSs and genomic features was computed. Given the vast heterogeneity of the annotations reported in the feature fields, a manual revision was required to fix potential inconsistencies in annotations, regarding both keywords and coordinates. A revision was performed when possible, while controversial and uncertain annotations were not considered. These analyses are presented as bar charts for individual viral classes and G-island pattern types (GG-, GGG-, GGGG-island) (S3–S5 Figs). As regards the GGG-island type, the herpesvirus family of dsDNA viruses presents PQSs distributed along all the four identified genomic features, with a particularly high concentration in RR and, in some members, in the 5'–regulatory region. This feature is consistent with the reported extent of G4s in HSV-1, which are mainly clustered in the RR of the virus genome [12, 13]. Conversely, viruses belonging to the ssRNA (+) and ssRNA (-) classes show PQSs mainly grouped in CDS and in the 3'- and 5'-regulatory regions, respectively. HIV-1, belonging to ssRNA (RT) virus class, presents PQSs of the GGG-island type mainly in the RR and 3'-regulatory regions and in part in CDS. This distribution confirms previous data [28, 32]. Conversely, other retroviruses (ssRNA (RT)) such as HTLV-1 and HTLV-2, display PQSs in the CDS. Given the lower stringency of PQSs of the GG-island type, these are more widely distributed along the four identified genomic features, whereas the most stringent PQSs of the GGGG type, present only in herpesviruses (dsDNA) and HTLV-1 (ssRNA (RT)), show a clear-cut localization in the RR and CDS, respectively. These data indicate that the localization of PQSs in the viral genomes differs in virus classes.

The number and type of PQSs are characteristic of virus classes

In this line of thinking, we asked whether the observed number of PQSs, and more precisely its statistical significance with respect to the two random assembling scenarios, is representative for a specific Baltimore class. To answer this question, we checked whether it is possible to classify each virus to one of the six classes considered, that is, dsDNA, ssDNA, dsRNA, ssRNA (+), ssRNA (-) and ssRNA (RT), based on the information of how significant its median PQS counts are. We used a classifier built on multinomial logistic regression, as this method is both interpretable and robust to unbalanced group sizes as long as the group sizes are large enough. To avoid the latter drawback, we excluded from the model fit the hepatitis B virus, the only virus classified as dsDNA (RT), and the two unclassified hepatitis delta and hepatitis E viruses. Six features were used to classify the viruses, i.e. the six mid-P values (those calculated for GG-, GGG-, GGGG-, both in the positive and negative strand) which qualify the PQS content of the real viral sequences. The values were multiplied by 1 or -1 depending on whether the median PQS count was over- or under-represented. Since real and corresponding simulated sequences contain the same base or G-islands composition, the classification model based on PQS content does not depend on the highly variable genome length and G/C content in the different virus classes but is specifically designed on the peculiar presence or absence of PQSs in each viral class. Furthermore, 34 viruses with no PQS count in all three G-island types in both the positive and negative strand and non-significant mid-P values at the 10% level were excluded from the analysis. We re-classified every viral genome used in our assessment using the discriminant function obtained from a leave-one-out analysis. This latter technique allowed us to accurately estimate how our classifier performs without the need to split our data into a training and a test set. The corresponding confusion matrix is given in Table 3 from where it is possible to extract the overall percentage of correct classifications that amount to 66.7% for the single nucleotide assembling model and 68.1% for the G-island reshuffling model. The agreement is good for the dsDNA, ssDNA, dsRNA, ssRNA (+) and ssRNA (-) classes. The two unclassified genomes of the hepatitis delta and hepatitis E viruses were classified as ssRNA (+).

Table 3. Confusion matrix for the semi-parametric classifier for G4 structure.

Single nucleotide assembling model		Predicted class						
		dsDNA	ssDNA	dsRNA	ssRNA (+)	ssRNA (-)	ssRNA (RT)	Unclassified
True class	dsDNA	33	0	0	1	4	1	0
	ssDNA	2	0	0	5	0	1	0
	dsRNA	0	0	13	0	5	0	0
	ssRNA (+)	4	0	0	45	13	0	0
	ssRNA (-)	3	0	8	18	48	0	0
	ssRNA (RT)	4	0	0	0	0	3	0
	Unclassified	0	0	0	2	0	0	0
G-island reshuffling model		**Predicted class**						
		dsDNA	ssDNA	dsRNA	ssRNA (+)	ssRNA (-)	ssRNA (RT)	Unclassified
True class	dsDNA	31	0	0	5	1	2	0
	ssDNA	4	0	0	2	2	0	0
	dsRNA	0	0	14	0	4	0	0
	ssRNA (+)	3	1	0	48	10	0	0
	ssRNA (-)	7	1	4	13	52	0	0
	ssRNA (RT)	3	0	0	3	1	0	0
	Unclassified	0	0	0	2	0	0	0

The number of viruses classified into the six Baltimore groups is shown, based on the two different simulation scenarios (single-nucleotide and G-island reshuffling). The classifier is based on a multinomial model and uses the one-sided mid-P values as features in combination with the information on whether the median PQS count is under- or over-represented.

https://doi.org/10.1371/journal.pcbi.1006675.t003

Discussion

In this work we provide: i) the list of PQSs present in all human virus genomes (both positive and negative strands), ii) their position in the viral genome, iii) the degree of conservation of both G-islands and loops vs. the genome, iv) the statistical significance of PQS abundance in each virus. Our data show that viruses belonging to dsDNA, ssDNA, ssRNA (RT) and, to a less extent, ssRNA (+) and ssRNA (-) display the largest presence of GG-, GGG- and GGGG-type PQSs (box plots, Fig 2) and that the presence of PQSs in all these virus groups is statistically significant (segment diagrams, S1 and S2 Figs). Both results support a role of G4s in the virus biology: indeed, some G4s predicted in this work were already reported in viruses and were shown to possess specific and different functions.

We evidenced some general trends and exceptions that are worth noting if seen in comparative terms among all viruses considered in this study. Starting from the general features we noted: i) high G/C content is not sufficient per se to generate a high number of PQSs, as observed in G/C rich members of *Herpesviridae* family that are richer of PQSs than expected. ii) The statistical significance of PQSs found in real sequences tends to decrease when G-islands reshuffling (ISL) is compared with the corresponding single nucleotide assembling (SN), as is appreciable from the heatmap in Fig 4 (more intense color in the heatmap boxes). This suggests that short sequences of a given length (k-mer) could be over- or under-represented in certain viral genomes, as already reported in the literature [41, 42]. We observed that viral genomes enriched in PQSs also contain a higher number of G-islands than expected from mere nucleotide composition, especially evident in the GG-islands. iii) The unevenly distribution of PQSs can be used to classify membership of a virus in its corresponding category. This was not predictable *a priori* but up to two-thirds of unequivocal assignments suggest that for some viruses the PQS content works as a distinctive feature. iv) PQS localization shows differences in some virus classes, but this outcome is still incomplete due to lack of information in the databases about virus genomic features and partitioning into regulatory and coding regions.

Some other interesting observations are worth reporting as either special cases or exceptions. To start with, the ssRNA (-) group is the most heterogeneous one, since some viral species are significantly enriched in PQSs up to the most extended G-island type (GGGG), while others lack this feature. Surprisingly, two viruses of the dsDNA group, which was generally highly enriched in PQSs, show a significantly lower presence of PQSs than expected in a random sequence with the same G/C content (SN, S3 Table), even though the opposite result was observed vs. simulated genomes with the same G-islands content as the real ones (ISL, S3 Table). These two viruses, i.e. *Molluscum contagiosum virus* and *Orf virus*, are the only ones belonging to genera other than the *Orthopoxvirus* within the *Poxviridae* family that cause skin lesions. Finally, dsRNA and dsDNA (RT) viruses are notably poor in PQSs and with mostly null statistical significance; however, single PQSs are highly conserved (e.g. rotavirus a segment 6), therefore still conveying potential biological interest.

These data indicate that PQSs are mainly present in ss-genome viruses, which in principle are more suitable to fold into G4s since they do not require unfolding from a fully complementary strand. The major exception to this evidence is the *Herpesviridae* family of dsDNA viruses. In this case, most PQSs reported here and also previously described [12, 13] form in repeated regions. It is possible that repeated sequences are more prone to alternative folding, as shown by several non-canonical structures that form in repeated regions of the DNA [46–49]. However, for some herpesviruses many PQSs are also present in regulatory regions, which may indicate yet undiscovered functional roles. To note that the investigation of PQSs was performed on a maximum window of 52 nucleotides in the case of isolated G4s. Alternatively, when more than four G-islands are found complying with the maximum distance allowed between consecutive islands, this window is extended as long as the rules are satisfied, thus including multiple distinct PQSs or potential isoforms. However, it is possible, especially for the ss genomes, that bases more distant in the primary sequence interact among each other, therefore expanding the repertoire of G4 structures.

The significant enrichment of PQSs in many viruses with respect to the corresponding randomized genomes is an indication that the clustering of G-islands did not occur by pure chance, suggesting a specific biological role of the G4 structures [40]. Complementary to this, the analysis of the PQS conservation highlights every PQS that is conserved among viral strains. Since one of the peculiarities of viral genomes is their fast mutation rate, the strong conservation of a specific G-island pattern among strains is per se an indication of the biological relevance of a PQS. In light of this, single conserved PQSs in viruses that do not display statistically significant PQS enrichment may retain key functional roles. The meaning of PQS conservation can have different explanations for the different viruses analyzed in this study. Given the high variability in the number of full-genome sequences available for each species, a more general evaluation of PQS conservation at higher taxonomic ranks (e.g. at the family level) could have been more informative. Nonetheless, generating and analyzing whole-genome multiple alignments involving different viral species, even if belonging to the same family, is almost prohibitive given the huge variability that is usually present in their genome sequences. Hence, the conservation of each PQS has to be considered on a case by case basis, exploiting the visualization tools provided in the website. As an example, an interesting approach could be looking at the discrepancy between the conservation of G-islands and connecting loops (*loop_conservation*) as an additional indication on the likeliness of biological implications of a specific PQS. A positive *loop_conservation* value highlights G-islands more conserved than their connecting loops, suggesting that only the PQS scaffold is required for mechanisms that are important for the virus life cycle, while the loops are dispensable. Considering the high mutation rate of viruses, this type of conservation indicates sequences where G4 formation is most likely essential. Equally conserved G-islands and loops (*loop_conservation* value = 0) imply that both the PQS scaffold and connecting loops are potentially relevant for the virus and probably involve interactors that specifically recognize them. In this case, the high sequence conservation, especially in CDS, may depend on the required conservation of that peculiar gene product rather than the presence of a G4 structure. Nonetheless, the option of targeting these conserved G4s for therapeutic purposes remains unaltered and the availability of specific and conserved loops may only enhance the chance of finding selective ligands [50]. Therefore, from this point of view, the 'zero' class is the best scenario for the development of specific drugs. The "negative" *loop_conservation* value scenario is of less immediate interpretation: it is possible that false positive hits fall in this category as it is unexpected that G-islands are less conserved than their connecting loops.

The evidence provided here, the previous studies on G4s in viruses, and the possibility to correctly classify the majority of viruses based on their PQSs (Table 3) suggest that most of the virus classes adopted G4-mediated mechanisms to control their viral cycles.

Together with the associated database, which is projected to be periodically updated to keep up with the fast-growing list of novel sequenced viruses, this work offers comprehensive data to guide researchers in the choice of the most significant PQSs within a human virus genome of interest. Hopefully, this will accelerate research in this area with the identification of new G4-mediated mechanisms in viruses and the development of effective and innovative therapeutics.

Materials and methods

PQS detection and evaluation of conservation

The complete list of viral species able to infect humans was retrieved from http://viralzone.expasy.org/all_by_species/678.html (accessed in April 2016) and, for each of them, all available complete genome sequences were downloaded from GenBank. Multiple alignments were built for each virus with usearch8 [51], using a permissive identity threshold (60%) to account for viral variability. Since in some cases nucleotide heterogeneity within viral species exceeded this value, multiple clusters of aligned sequences were obtained for some viruses, representing distinct genotypes. Considering the difficulty of obtaining high quality alignments beyond this limit of nucleotide similarity, all clusters obtained with this method were kept separate, manually assigned to specific genotypes and independently processed in the downstream analyses. One genome per each group of aligned sequences was chosen to serve as reference sequence, possibly belonging to the manually curated RefSeq database [https://www.ncbi.nlm.nih.gov/refseq/]. The complete list of selected reference sequences is reported in S1 Table.

PQSs were searched in all multiple-aligned nucleotide sequences with an in-house developed tool, as previously described [36, 52]. Briefly, a PQS was reported when at least four consecutive guanine islands (G-islands) were detected. If '*l*' is the minimum number of G in every G-island of a PQS and '*d*' is the maximum distance allowed between two consecutive G-islands, the following combinations of '*l*' and '*d*' were searched: $l = 2$ and $d = 7$; $l = 3$ and $d = 12$; $l = 4$ and $d = 12$. Patterns in the negative strands of viral genomes were searched by looking for cytosines (Cs) instead of Gs. The conservation of each PQS in the multiple aligned genomes of the viruses was determined by looking at the conservation not only of the G-islands but also their connecting loops. We computed different indexes to measure the nucleotide sequence conservation of viral genomes and PQSs:

1. *G4_scaffold_conservation_index* (*G4_SCI*): it is referred to the G-islands. For each virus and for every detected PQS, it is calculated as the percentage of independent genomes maintaining the corresponding G-islands:

$$G4\_SCI = \frac{N_{G\_islands}}{N_{tot}} * 100$$

   where $N_{G\_islands}$ is the number of sequences possessing the G-islands in a certain genome position and N_{tot} is the total number of sequences available for the virus.

2. *Loop_conservation*: it is the difference between G4_SCI and the local conservation of the viral sequence spanning the PQS (LSC_{G4}).

$$Loop\_conservation = G4\_SCI - LSC_{G4}$$

LSC_{G4} is calculated as the average of LSC windows overlapping the PQS. LSC measure is computed within a sliding window of fixed length (length 20, shift 10), averaging the conservation values of each position extracted from the multiple sequence alignments with Jalview [53]. They are formally defined as:

$$LSC = \frac{\sum_{i=1}^{20} c_{\max i}}{20}$$

$$LSC_{G4} = \frac{LSC_1 + \ldots + LSC_n}{n}$$

where $c_{max\ i}$ is the maximum conservation at position i of the multiple aligned sequences and n is the number of windows overlapping the PQS.The results of these analyses are presented individually for each virus and PQS (http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=g4virus), together with the calculated profiles of simple linguistic complexity and Shannon entropy that can highlight other potential local features of the sequence (e.g. repeats and low complexity regions) [54]. All charts were generated with Plotly [https://plot.ly], exploiting Pandas [55] and Numpy Python libraries [56]. Multiple alignments are visualized with MSAViewer[57] and genomic features with JBrowse 1.15.0 [58]. Unless otherwise stated, analyses were conducted with ad hoc developed Python and Perl scripts, which are available in the website (scripts.tar.gz).

Evaluation of PQS conservation in real vs randomized viral sequences

To determine whether the presence of PQSs in a virus is a conserved feature or it is only a consequence of its nucleotide composition, simulated viral genomes were generated and compared with real data. Two different strategies were adopted to generate simulated data:

1. *Single nucleotide assembling (SN)*. A computational approach was adopted where, in analogy to Huppert and Balasubramanian [43], the viral genome was modelled as a multinomial stream based on the assumption that each DNA base is independent. These authors give an explicit solution for the prevalence of PQSs in the human genome as a function of *p(G)*, the probability of any base being G. In our approach, we also accounted for the probability of cytosines (*p(C)*) and additionally assumed that adenine (A) and thymine (T) bases were equally likely to occur. As all four probabilities need to sum up to one, the statistical reference model is a multinomial distribution with probability vector *(p(G), p(C), p(A), p(T))*. We hence took as many independent draws from this multinomial distribution as the number of nucleotides in the reference viral genome (S1 Table). The probabilities *p(G)* and *p(C)*

vary for each virus and reflect the prevalence of G and C bases present in that virus, while the remaining proportion is equally split to give *p(A)* and *p(T)*. For each virus, 10,000 independent sequences were produced *in silico* with this method; the 'sample' R command with replacement was used and the provided parameters were the genome length and the relative abundance of the four nucleotides in the real genomes.

2. *G-islands reshuffling (ISL)*. For each virus, we generated a scaffold of length *n* made of only As, with *n* corresponding to the length of the virus reference genome; then, we replaced di-, tri-, or tetra-nucleotides, at random positions and without overlaps, with as many GG-, GGG-, GGGG-, CC-, CCC-, CCCC-islands as in the reference sequence, respectively. Overall, we generated 10,000 independent sequences for three different simulated datasets, one for each island length.

Statistical methods

The simulated sequences were scanned for the presence of PQSs as previously described. The 10,000 counts obtained for each simulation formed the empirical distribution for PQS prevalence under the hypothesis of random assembling of the genome in the SN and ISL models respectively. The mid-P value was calculated using a homemade function. The semiparametric classifier used to assign the virus to its exact class relying on its PQS content was based on the 'multinom' function of the R package 'nnet'.

PQSs position and overlap with genomic features

The feature tables containing viral genome annotations were downloaded from RefSeq or GenBank for all the reference sequences reported in S1 Table. Genome coordinates were extracted for coding sequences ('CDS'), repeat regions ('repeat_region'), 5'- and 3'-untranslated (UTR) and promoter regions. Given the annotation inconsistency of promoters and UTRs, two new feature categories were created, 5'–and 3'–regulatory regions that were defined by exploiting the annotation of genes and CDSs. We calculated boundaries for genes in the positive strand of viral genomes as follows: 5'–regulatory = $S_{gene} - S_{cds}$; 3'–regulatory = E_{cds}—E_{gene}. For the genes in the negative strand of viral genomes we defined: 5'–regulatory = Scds—S_{gene}; 3'–regulatory = E_{gene}—E_{cds}. S_{gene}, S_{cds}, E_{gene} and E_{cds} are the Start (S) and End (E) of genes and CDSs as extracted from the feature tables. These newly defined features contain both UTRs and promoters. Since the positive genomic strands are deposited in RefSeq for most of the viruses belonging to the ssRNA (-) class, the following sequences available as negative strands were converted into their inverse complement, together with the coordinates of their genomic features: Junin arenavirus segment S (NC_005081) and segment L (NC_005080), Lassa virus segment L (NC_004297), lymphocytic choriomeningitis virus segment S (GQ862982), Machupo virus segment S (AY924208) and L (AY624354), Pichinde virus segment S (NC_006447), Rift Valley fever virus segment S (NC_014395), and Toscana virus segment S (NC_006318). The overlap extent between PQSs and genomic features was computed by intersecting the genomic coordinates of each PQS with the genomic features extracted from the corresponding virus, and a positive count was recorded every time an overlap of at least one nucleotide was detected. Finally, to compare the enrichment in different feature classes, characterized by different sizes, a normalization step was performed. The total extension of each feature class (*i.e.* CDS, repeat_region, 5'–regulatory and 3'–regulatory) was calculated by summing the lengths of individual features. The total count of PQS overlapping a feature class was then divided by the total length of the class and multiplied by a factor 1,000 to obtain the number of PQS present every 1,000 nucleotides. This procedure was performed considering only the PQSs conserved in at least 80% of sequences for each viral species. All feature tables files were manually revised to fix inconsistencies in the use of keywords and coordinates.

Supporting information

S1 Fig. PQS content in real vs simulated viral genomes (single nucleotide assembling). Segment diagrams of mid-P values obtained by comparing the PQS content detected in real and simulated viral genomes. Simulated viruses have the same nucleotide composition of the real ones, but different order. The three G-island types considered in the positive (+) and negative (-) strands of all human virus genomes are grouped in the 7 Baltimore classes. From left to right, each segment represents one of the three G-islands (GG, GGG, GGGG) in the positive (top half) and negative (bottom half) strands; the radius of a segment corresponds to 1 minus the mid-P value. Thus, full segments indicate highly significant PQSs, whereas null segments indicate non-significant PQSs, with respect to the random sequences.
(TIF)

S2 Fig. PQS content in real vs simulated viral genomes (G-island reshuffling). Segment diagrams of mid-P values obtained by comparing the PQS content detected in real and simulated viral genomes. Simulated viruses are obtained by reshuffling the positions of their GG-, GGG- or GGGG-islands. The three G-island types considered in the positive (+) and negative (-) strands of all human virus genomes are grouped in the 7 Baltimore classes. From left to right, each segment represents one of the three G-islands (GG, GGG, GGGG) in the positive (top half) and negative (bottom half) strands; the radius of a segment corresponds to 1 minus the mid-P value. Thus, full segments indicate highly significant PQSs, whereas null segments indicate non-significant PQSs, with respect to the random sequences.
(TIF)

S3 Fig. PQS overlap with genomic features—GG islands. Representative figure of GG-island PQSs that overlap with genomic features. The bar charts report the distribution of PQSs in genomic features where available and annotated in the database. The number of PQSs per 1kb is reported on the x-axis both for the positive (orange) and negative (blue) strands. The four features considered are coding sequences (CDS), repeat regions (RR), and regulatory regions at the 5' and 3' ends.
(TIF)

S4 Fig. PQS overlap with genomic features—GGG islands. Representative figure of GGG-island PQSs that overlap with genomic features. The bar charts report the distribution of PQSs in genomic features where available and annotated in the database. The number of PQSs per 1kb is reported on the x-axis both for the positive (orange) and negative (blue) strands. The four features considered are coding sequences (CDS), repeat regions (RR), and regulatory regions at the 5' and 3' ends.
(TIF)

S5 Fig. PQS overlap with genomic features—GGGG islands. Representative figure of GGGG-island PQSs that overlap with genomic features. The bar charts report the distribution of PQSs in genomic features where available and annotated in the database. The number of PQSs per 1kb is reported on the x-axis both for the positive (orange) and negative (blue) strands. The four features considered are coding sequences (CDS), repeat regions (RR), and

regulatory regions at the 5' and 3' ends.
(TIF)

S1 Table. Accession numbers of reference sequences selected for each virus.
(DOCX)

S2 Table. Experimentally validated G4s in human viruses.
(DOCX)

S3 Table. List of viruses whose PQS content is significant at 10% with respect to randomized sequences.
(XLSX)

Author Contributions

Conceptualization: Sara N. Richter, Stefano Toppo.

Data curation: Enrico Lavezzo, Michele Berselli.

Formal analysis: Ilaria Frasson, Alessandra R. Brazzale.

Funding acquisition: Sara N. Richter.

Investigation: Enrico Lavezzo, Michele Berselli, Ilaria Frasson, Rosalba Perrone.

Methodology: Enrico Lavezzo, Michele Berselli, Alessandra R. Brazzale.

Project administration: Stefano Toppo.

Resources: Stefano Toppo.

Software: Enrico Lavezzo, Michele Berselli.

Supervision: Giorgio Palù, Alessandra R. Brazzale, Sara N. Richter, Stefano Toppo.

Validation: Ilaria Frasson, Alessandra R. Brazzale.

Visualization: Michele Berselli.

Writing – original draft: Enrico Lavezzo, Stefano Toppo.

Writing – review & editing: Sara N. Richter, Stefano Toppo.

References

1. Lipps HJ, Rhodes D. G-quadruplex structures: in vivo evidence and function. Trends Cell Biol. 2009; 19(8):414–22. https://doi.org/10.1016/j.tcb.2009.05.002 PMID: 19589679.
2. Sen D, Gilbert W. A sodium-potassium switch in the formation of four-stranded G4-DNA. Nature. 1990; 344(6265):410–4. https://doi.org/10.1038/344410a0 PMID: 2320109.
3. Maizels N, Gray LT. The G4 genome. PLoS Genet. 2013; 9(4):e1003468. https://doi.org/10.1371/journal.pgen.1003468 PMID: 23637633.
4. Rouleau S, Jodoin R, Garant JM, Perreault JP. RNA G-Quadruplexes as Key Motifs of the Transcriptome. Adv Biochem Eng Biotechnol. 2017. https://doi.org/10.1007/10_2017_8 PMID: 28382477.
5. Jayaraj GG, Pandey S, Scaria V, Maiti S. Potential G-quadruplexes in the human long non-coding transcriptome. RNA Biol. 2012; 9(1):81–6. https://doi.org/10.4161/rna.9.1.18047 PMID: 22258148.
6. Mirihana Arachchilage G, Dassanayake AC, Basu S. A potassium ion-dependent RNA structural switch regulates human pre-miRNA 92b maturation. Chem Biol. 2015; 22(2):262–72. https://doi.org/10.1016/j.chembiol.2014.12.013 PMID: 25641166.
7. Agarwala P, Pandey S, Maiti S. The tale of RNA G-quadruplex. Org Biomol Chem. 2015; 13(20):5570–85. https://doi.org/10.1039/c4ob02681k PMID: 25879384.

8. Cammas A, Millevoi S. RNA G-quadruplexes: emerging mechanisms in disease. Nucleic Acids Res. 2017; 45(4):1584–95. https://doi.org/10.1093/nar/gkw1280 PMID: 28013268.

9. Flint SJ, Racaniello VR, Glenn FR, Skalka AM, Enquist LW. Principles of Virology: Volume 1 Molecular Biology. 4th ed: ASM Press; 2015 August 17, 2015. 574 p.

10. Metifiot M, Amrane S, Litvak S, Andreola ML. G-quadruplexes in viruses: function and potential therapeutic applications. Nucleic Acids Res. 2014; 42(20):12352–66. Epub 2014/10/22. https://doi.org/10.1093/nar/gku999 PMID: 25332402.

11. Ruggiero E, Richter SN. G-quadruplexes and G-quadruplex ligands: targets and tools in antiviral therapy. Nucleic Acids Res. 2018. https://doi.org/10.1093/nar/gky187 PMID: 29554280.

12. Artusi S, Nadai M, Perrone R, Biasolo MA, Palu G, Flamand L, et al. The Herpes Simplex Virus-1 genome contains multiple clusters of repeated G-quadruplex: Implications for the antiviral activity of a G-quadruplex ligand. Antiviral Res. 2015; 118:123–31. Epub 2015/04/07. https://doi.org/10.1016/j.antiviral.2015.03.016 PMID: 25843424.

13. Artusi S, Perrone R, Lago S, Raffa P, Di Iorio E, Palu G, et al. Visualization of DNA G-quadruplexes in herpes simplex virus 1-infected cells. Nucleic Acids Res. 2016; 44(21):10343–53. https://doi.org/10.1093/nar/gkw968 PMID: 27794039.

14. Gilbert-Girard S, Gravel A, Artusi S, Richter SN, Wallaschek N, Kaufer BB, et al. Stabilization of Telomere G-Quadruplexes Interferes with Human Herpesvirus 6A Chromosomal Integration. J Virol. 2017; 91(14). https://doi.org/10.1128/JVI.00402-17 PMID: 28468887.

15. Madireddy A, Purushothaman P, Loosbroock CP, Robertson ES, Schildkraut CL, Verma SC. G-quadruplex-interacting compounds alter latent DNA replication and episomal persistence of KSHV. Nucleic Acids Res. 2016; 44(8):3675–94. https://doi.org/10.1093/nar/gkw038 PMID: 26837574.

16. Murat P, Zhong J, Lekieffre L, Cowieson NP, Clancy JL, Preiss T, et al. G-quadruplexes regulate Epstein-Barr virus-encoded nuclear antigen 1 mRNA translation. Nat Chem Biol. 2014; 10(5):358–64. Epub 2014/03/19. https://doi.org/10.1038/nchembio.1479 PMID: 24633353.

17. Norseen J, Johnson FB, Lieberman PM. Role for G-quadruplex RNA binding by Epstein-Barr virus nuclear antigen 1 in DNA replication and metaphase chromosome attachment. J Virol. 2009; 83 (20):10336–46. Epub 2009/08/07. https://doi.org/10.1128/JVI.00747-09 PMID: 19656898.

18. Tellam JT, Zhong J, Lekieffre L, Bhat P, Martinez M, Croft NP, et al. mRNA Structural constraints on EBNA1 synthesis impact on in vivo antigen presentation and early priming of CD8+ T cells. PLoS Pathog. 2014; 10(10):e1004423. https://doi.org/10.1371/journal.ppat.1004423 PMID: 25299404.

19. Lista MJ, Martins RP, Billant O, Contesse MA, Findakly S, Pochard P, et al. Nucleolin directly mediates Epstein-Barr virus immune evasion through binding to G-quadruplexes of EBNA1 mRNA. Nat Commun. 2017; 8:16043. https://doi.org/10.1038/ncomms16043 PMID: 28685753.

20. Tluckova K, Marusic M, Tothova P, Bauer L, Sket P, Plavec J, et al. Human papillomavirus G-quadruplexes. Biochemistry. 2013; 52(41):7207–16. Epub 2013/09/21. https://doi.org/10.1021/bi400897g PMID: 24044463.

21. Satkunanathan S, Thorpe R, Zhao Y. The function of DNA binding protein nucleophosmin in AAV replication. Virology. 2017; 510:46–54. https://doi.org/10.1016/j.virol.2017.07.007 PMID: 28704696.

22. Fleming AM, Ding Y, Alenko A, Burrows CJ. Zika Virus Genomic RNA Possesses Conserved G-Quadruplexes Characteristic of the Flaviviridae Family. ACS Infect Dis. 2016; 2(10):674–81. https://doi.org/10.1021/acsinfecdis.6b00109 PMID: 27737553.

23. Wang SR, Min YQ, Wang JQ, Liu CX, Fu BS, Wu F, et al. A highly conserved G-rich consensus sequence in hepatitis C virus core gene represents a new anti-hepatitis C target. Sci Adv. 2016; 2(4): e1501535. Epub 2016/04/07. https://doi.org/10.1126/sciadv.1501535 PMID: 27051880.

24. Tan J, Vonrhein C, Smart OS, Bricogne G, Bollati M, Kusov Y, et al. The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. PLoS Pathog. 2009; 5(5): e1000428. Epub 2009/05/14. https://doi.org/10.1371/journal.ppat.1000428 PMID: 19436709.

25. Kusov Y, Tan J, Alvarez E, Enjuanes L, Hilgenfeld R. A G-quadruplex-binding macrodomain within the "SARS-unique domain" is essential for the activity of the SARS-coronavirus replication-transcription complex. Virology. 2015; 484:313–22. https://doi.org/10.1016/j.virol.2015.06.016 PMID: 26149721.

26. Wang SR, Zhang QY, Wang JQ, Ge XY, Song YY, Wang YF, et al. Chemical Targeting of a G-Quadruplex RNA in the Ebola Virus L Gene. Cell Chem Biol. 2016; 23(9):1113–22. https://doi.org/10.1016/j.chembiol.2016.07.019 PMID: 27617851.

27. Biswas B, Kandpal M, Vivekanandan P. A G-quadruplex motif in an envelope gene promoter regulates transcription and virion secretion in HBV genotype B. Nucleic Acids Res. 2017. https://doi.org/10.1093/nar/gkx823 PMID: 28981800.

28. Perrone R, Nadai M, Frasson I, Poe JA, Butovskaya E, Smithgall TE, et al. A dynamic G-quadruplex region regulates the HIV-1 long terminal repeat promoter. J Med Chem. 2013; 56(16):6521–30. Epub 2013/07/20. https://doi.org/10.1021/jm400914r PMID: 23865750.

29. Perrone R, Butovskaya E, Daelemans D, Palu G, Pannecouque C, Richter SN. Anti-HIV-1 activity of the G-quadruplex ligand BRACO-19. J Antimicrob Chemother. 2014; 69(12):3248–58. Epub 2014/08/12. https://doi.org/10.1093/jac/dku280 PMID: 25103489.

30. Piekna-Przybylska D, Sullivan MA, Sharma G, Bambara RA. U3 region in the HIV-1 genome adopts a G-quadruplex structure in its RNA and DNA sequence. Biochemistry. 2014; 53(16):2581–93. Epub 2014/04/17. https://doi.org/10.1021/bi4016692 PMID: 24735378.

31. Amrane S, Kerkour A, Bedrat A, Vialet B, Andreola ML, Mergny JL. Topology of a DNA G-quadruplex structure formed in the HIV-1 promoter: a potential target for anti-HIV drug development. J Am Chem Soc. 2014; 136(14):5249–52. Epub 2014/03/22. https://doi.org/10.1021/ja501500c PMID: 24649937.

32. Perrone R, Nadai M, Poe JA, Frasson I, Palumbo M, Palu G, et al. Formation of a unique cluster of G-quadruplex structures in the HIV-1 Nef coding region: implications for antiviral activity. PLoS One. 2013; 8(8):e73121. https://doi.org/10.1371/journal.pone.0073121 PMID: 24015290.

33. Lago S, Tosoni E, Nadai M, Palumbo M, Richter SN. The cellular protein nucleolin preferentially binds long-looped G-quadruplex nucleic acids. Biochim Biophys Acta. 2017; 1861(5 Pt B):1371–81. https://doi.org/10.1016/j.bbagen.2016.11.036 PMID: 27913192.

34. Scalabrin M, Frasson I, Ruggiero E, Perrone R, Tosoni E, Lago S, et al. The cellular protein hnRNP A2/B1 enhances HIV-1 transcription by unfolding LTR promoter G-quadruplexes. Sci Rep. 2017; 7:45244. https://doi.org/10.1038/srep45244 PMID: 28338097.

35. Perrone R, Doria F, Butovskaya E, Frasson I, Botti S, Scalabrin M, et al. Synthesis, Binding and Antiviral Properties of Potent Core-Extended Naphthalene Diimides Targeting the HIV-1 Long Terminal Repeat Promoter G-Quadruplexes. J Med Chem. 2015; 58(24):9639–52. Epub 2015/11/26. https://doi.org/10.1021/acs.jmedchem.5b01283 PMID: 26599611.

36. Perrone R, Lavezzo E, Palu G, Richter SN. Conserved presence of G-quadruplex forming sequences in the Long Terminal Repeat Promoter of Lentiviruses. Sci Rep. 2017; 7(1):2018. https://doi.org/10.1038/s41598-017-02291-1 PMID: 28515481.

37. Pandey S, Agarwala P, Maiti S. Effect of loops and G-quartets on the stability of RNA G-quadruplexes. J Phys Chem B. 2013; 117(23):6896–905. https://doi.org/10.1021/jp401739m PMID: 23683360.

38. Guedin A, Gros J, Alberti P, Mergny JL. How long is too long? Effects of loop size on G-quadruplex stability. Nucleic Acids Res. 2010; 38(21):7858–68. https://doi.org/10.1093/nar/gkq639 PMID: 20660477.

39. Mukundan VT, Phan AT. Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. J Am Chem Soc. 2013; 135(13):5017–28. https://doi.org/10.1021/ja310251r PMID: 23521617.

40. Frees S, Menendez C, Crum M, Bagga PS. QGRS-Conserve: a computational method for discovering evolutionarily conserved G-quadruplex motifs. Hum Genomics. 2014; 8:8. https://doi.org/10.1186/1479-7364-8-8 PMID: 24885782.

41. Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. 2008; 4(6):e1000079. https://doi.org/10.1371/journal.ppat.1000079 PMID: 18535658.

42. Lobo FP, Mota BE, Pena SD, Azevedo V, Macedo AM, Tauch A, et al. Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts. PLoS One. 2009; 4(7):e6282. https://doi.org/10.1371/journal.pone.0006282 PMID: 19617912.

43. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic Acids Res. 2005; 33(9):2908–16. https://doi.org/10.1093/nar/gki609 PMID: 15914667.

44. Armitage P, Berry G. Statistical Methods in Medical Research. Publications OBS, editor 1994.

45. Chambers J, Cleveland W, Kleiner B, Tukey P. Graphical Methods for Data Analysis. Wadsworth, editor 1983.

46. Fry M, Loeb LA. The fragile X syndrome d(CGG)n nucleotide repeats form a stable tetrahelical structure. Proc Natl Acad Sci U S A. 1994; 91(11):4950–4. Epub 1994/05/24. PMID: 8197163.

47. Sket P, Pohleven J, Kovanda A, Stalekar M, Zupunski V, Zalar M, et al. Characterization of DNA G-quadruplex species forming from C9ORF72 G4C2-expanded repeats associated with amyotrophic lateral sclerosis and frontotemporal lobar degeneration. Neurobiol Aging. 2015; 36(2):1091–6. Epub 2014/12/03. https://doi.org/10.1016/j.neurobiolaging.2014.09.012 PMID: 25442110.

48. Zhou B, Liu C, Geng Y, Zhu G. Topology of a G-quadruplex DNA formed by C9orf72 hexanucleotide repeats associated with ALS and FTD. Sci Rep. 2015; 5:16673. Epub 2015/11/14. https://doi.org/10.1038/srep16673 PMID: 26564809.

49. Reddy K, Zamiri B, Stanley SY, Macgregor RB Jr., Pearson CE. The disease-associated r(GGGGCC)n repeat from the C9orf72 gene forms tract length-dependent uni- and multimolecular RNA G-quadruplex structures. J Biol Chem. 2013; 288(14):9860–6. https://doi.org/10.1074/jbc.C113.452532 PMID: 23423380.

50. Parrotta L, Ortuso F, Moraca F, Rocca R, Costa G, Alcaro S, et al. Targeting unimolecular G-quadruplex nucleic acids: a new paradigm for the drug discovery? Expert Opin Drug Discov. 2014; 9(10):1167–87. https://doi.org/10.1517/17460441.2014.941353 PMID: 25109710.

51. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26(19):2460–1. https://doi.org/10.1093/bioinformatics/btq461 PMID: 20709691

52. Perrone R, Lavezzo E, Riello E, Manganelli R, Palu G, Toppo S, et al. Mapping and characterization of G-quadruplexes in Mycobacterium tuberculosis gene promoter regions. Sci Rep. 2017; 7(1):5743. https://doi.org/10.1038/s41598-017-05867-z PMID: 28720801.

53. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009; 25(9):1189–91. https://doi.org/10.1093/bioinformatics/btp033

54. Berselli M, Lavezzo E, Toppo S. NeSSie: a tool for the identification of approximate DNA sequence symmetries. Bioinformatics. 2018. https://doi.org/10.1093/bioinformatics/bty142 PMID: 29522153.

55. McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. 2010:51–6.

56. van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering. 2011; 13(2):22–30.

57. Yachdav G, Wilzbach S, Rauscher B, Sheridan R, Sillitoe I, Procter J, et al. MSAViewer: interactive JavaScript visualization of multiple sequence alignments. Bioinformatics. 2016; 32(22):3501–3. https://doi.org/10.1093/bioinformatics/btw474 PMID: 27412096.

58. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. Genome Res. 2009; 19(9):1630–8. https://doi.org/10.1101/gr.094607.109 PMID: 19570905.