DNA Research, 2020, 27(2), 1–16
doi: 10.1093/dnares/dsaa008
Advance Access Publication Date: 27 April 2020
Full Paper

OXFORD

Full Paper

Evolutionary selection against short nucleotide sequences in viruses and their related hosts

Yoram Zarai[1], Zohar Zafrir[1,2], Bunpote Siridechadilok[3], Amporn Suphatrakul[3], Modi Roopin[1,2], Justin Julander[4], and Tamir Tuller [1,2]*

[1]Biomedical Engineering Department, Tel Aviv University, Tel Aviv 69978, Israel, [2]SynVaccine Ltd., Ramat Hachayal, Tel Aviv, Israel, [3]National Center for Genetic Engineering and Biotechnology, Pathumthani 12120, Thailand, and [4]Institute for Antiviral Research, Utah State University, Logan, UT, USA

*To whom correspondence should be addressed. Tel. +11 972 3 6405836. Fax. +11 972 3 6405836.
Email: tamirtul@tauex.tau.ac.il

Received 2 January 2020; Editorial decision 17 April 2020; Accepted 20 April 2020

Abstract

Viruses are under constant evolutionary pressure to effectively interact with the host intracellular factors, while evading its immune system. Understanding how viruses co-evolve with their hosts is a fundamental topic in molecular evolution and may also aid in developing novel viral based applications such as vaccines, oncologic therapies, and anti-bacterial treatments. Here, based on a novel statistical framework and a large-scale genomic analysis of 2,625 viruses from all classes infecting 439 host organisms from all kingdoms of life, we identify short nucleotide sequences that are under-represented in the coding regions of viruses and their hosts. These sequences cannot be explained by the coding regions' amino acid content, codon, and dinucleotide frequencies. We specifically show that short homooligonucleotide and palindromic sequences tend to be under-represented in many viruses probably due to their effect on gene expression regulation and the interaction with the host immune system. In addition, we show that more sequences tend to be under-represented in dsDNA viruses than in other viral groups. Finally, we demonstrate, based on *in vitro* and *in vivo* experiments, how under-represented sequences can be used to attenuated Zika virus strains.

Key words: systems-biology, under-represented sequences, virus–host co-evolution, Zika virus

1 Introduction

Viruses, the most abundant type of biological entity, are small infectious agents that can only replicate inside the living cells of other organisms (hosts).[1] The viral genetic material is composed of either RNA or DNA molecule, single or double stranded. Viral genomes typically encode three types of protein: proteins for replicating the genome, proteins for packing the genome, and proteins for modifying the function of the host's cell to enhance the replication of the virus's material.[2,3]

Viruses are believed to play a central role in evolution,[4] (e.g. via horizontal gene transfer[2,5–7]), be responsible for various human diseases (e.g. AIDS and respiratory diseases[8,9]), and also have important applications to biotechnology[10] and nanotechnology.[11] For instance, the recent Zika virus (ZIKV) epidemic in the Americas have led the World Health Organization to declare a 'public health emergency of international concern',[12,13] and just recently the novel coronavirus (2019-nCoV) outbreak in China was declared pandemic by the same organization.[14] Due to their complete reliance on the host gene expression machinery, viruses are under constant evolutionary pressure to effectively interact with the host intracellular factors, and at the same time effectively evade its immune system.[3,15] Thus, understanding how viruses co-evolve with their hosts to ensure their fitness may help in developing novel viral based applications such as vaccines, oncologic therapies, and anti-bacterial treatments.

© The Author(s) 2020. Published by Oxford University Press on behalf of Kazusa DNA Research Institute.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

It is natural to expect that viruses and hosts co-evolution patterns are also encrypted in the viral genome. For example, it was shown that high correlation of GC content exists between bacteriophage and related hosts,[16] that a pattern of CpG dinucleotides is suppressed in vertebrate hosts and in their related RNA viruses,[17,18] that the frequency of TpA dinucleotides is suppressed in invertebrate hosts and in their related RNA viruses,[19] and that many long sequences are shared between hosts and their related viruses.[20,21]

Identification and analysis of short DNA sequences that are under-represented (also referred to as suppressed or avoided) in genomes of different species were analysed in the past.[22,33] For example, in,[22] Markov chain models were used to analyse short sequences in the DNA of two hosts: *Escherichia coli* and *Bacillus subtilis*. Markovian models were used in[23] to predict the frequencies of short sequences and applied them to many prokaryotic species, and the authors in[24] introduced an efficient algorithm to identify sequences that are avoided.

In this paper, we analyse under-represented nucleotide sequences in the coding regions of all types of viruses and in the coding regions of their corresponding hosts using a novel statistical framework. These sequences are analysed separately in each of the three reading frames. We provide a large database of these sequences, identify unique and interesting patterns within these sequences, and demonstrate how these sequences can be utilized to attenuate the ZIKV via *in vitro* and *in vivo* experiments.

2 Materials and methods

In this section, we briefly describe the main steps of our methodology. A detailed description appears in the Supplementary document.

2.1 Analysis flow overview

The general flow of our analysis is depicted in Fig. 1A. The dataset of virus–host associations was retrieved from previously published data.[34] These include 2,625 unique viruses and 439 corresponding hosts, where all the corresponding coding sequences were downloaded and processed. Randomization models were used to generate many random variants of the host and virus coding sequences. Two different randomization models were used, each control for different biases. A dinucleotide randomization model preserves both amino-acid order and content and the distribution of all 16 possible pairs of nucleotides, whereas a synonymous codon randomization model preserves both amino-acid order and content, and the codon usage bias. These were then used to statistically infer short nucleotide sequences that are under-represented within both the original host and virus genome coding regions, in each reading frame, and those that are common to all three reading frames. These under-represented sequences were analysed and compared among different viral groups and viral proteins, revealing some interesting evolutionary patterns that will be discussed later on. Based on this analysis, an attenuated variant of the ZIKV was engineered and its attenuation was demonstrated in cell lines and in mice.

2.2 Database

The virus and host coding sequences and association information was retrieved from a published database.[21] In brief, the association between viruses and hosts was derived from the GenomeNet Virus-Host Database.[34] The database contains 2,625 unique viruses and 439 corresponding unique hosts from all kingdoms of life (see Supplementary Table S1). Figure 1B depicts the six host domains in the database (vertebrates, bacteria, fungi, metazoa, planta, and

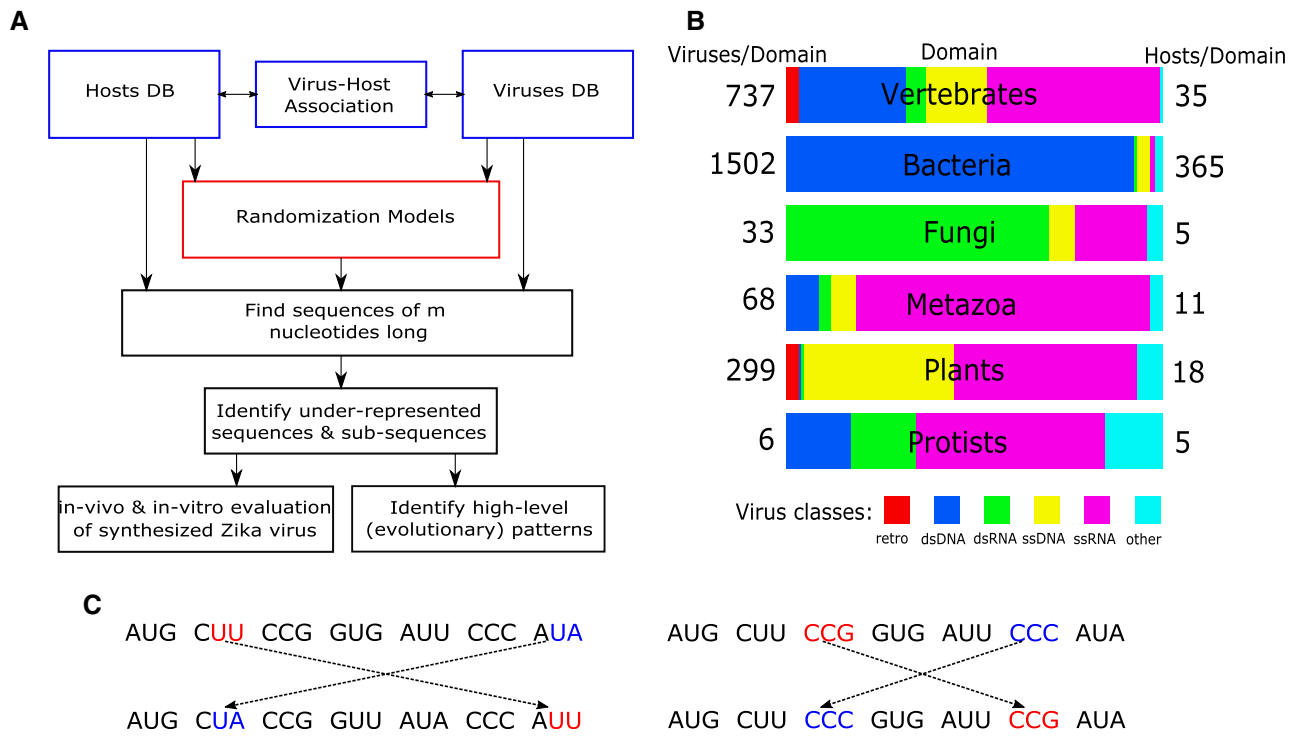

Figure 1. The analysis flow diagram (A), a summary of the viruses–hosts association database (B), where left values specify the total number of viruses corresponding to each host domain, and right values specify the total number of hosts in each host domain, and the randomization models (C), illustrating an example of dinucleotides randomization (left) and synonymous codons randomization (right).

protists), where we specify for each host domain the portion of the corresponding viruses belonging to each virus type. The virus types in the database are reverse-transcribing (retro), double-stranded DNA (dsDNA), double-stranded RNA (dsRNA), single-stranded DNA (ssDNA), single-stranded RNA (ssRNA, positive and negative sense), and other (unclassified).

2.3 Randomization models and statistical analysis

The question that we must first address is: what constitutes an under-represented sequence in a coding region? To detect sequences that are statistically under-represented in the coding regions, our statistical background model must capture well-understood coding region features, which are known to be under selection. For example, selection for codon usage bias may cause few short sequences to be in low abundance in the coding regions (as opposed, for example, to regions that are not translated). This, however, does not imply that these short sequences were directly selected against by evolutionary forces. Our definition of under-represented short nucleotide sequences in the coding region must then be formulated with respect to all known coding region features (i.e. amino-acids content and order, codon usage bias, and dinucleotide distribution), to suggest possibly new evolutionary forces acting on the viral coding regions.

To that end, two randomization models were used to evaluate our hypothesis for short, under-represented nucleotide sequences in the coding regions of the viruses and in the coding regions of their corresponding hosts. The first, called dinucleotide randomization, preserves both amino acid order and content (and thus the resulting protein), and the frequencies of the 16 possible pairs of adjacent nucleotides (dinucleotides). The second, called synonymous codon randomization preserves both amino-acids order and content (and thus the resulting protein) and the codon usage bias. Figure 1C depicts a schematic description of both randomization methods.

A selection against short nucleotide sequences that cannot be explained by the canonical genomic features that are preserved by both randomization models implies that these sequences will appear more frequently in the random variants (generated by the above randomization models) than in the original genome. Empirical *P*-values were derived from the empirical null model defined by the above two randomization models. The *P*-value estimates the probability of obtaining a random value (i.e. the number of occurrences of a sequence in the coding regions) that is the same or larger than the observed value in the original genome. This was performed separately in each of the three reading frames. A sequence was declared under-represented if its *P*-values corresponding to the two randomization models were both ≤ 0.05. Note that in the case of synonymous codon randomization, no under-represented sequence of size three nucleotides can be identified in the first reading frame.

Specifically, when analysing under-represented sequences in the viruses, we compared the original genome to 1,000 corresponding randomization variants generated by each of the randomization models described above. Under-represented sequences were then identified separately in each reading frame. In addition, common under-represented nucleotide sequences were identified (i.e. sequences that are under-represented in all three reading frames—see Supplementary document, Section 1.4.1). This may indicate selection against sequences that may 'interfere' with the process of mRNA translation. See Supplementary document, Sections 1.4 and 2.3 for an additional method of identifying under-represented sequences in the viruses based on the corresponding hosts (i.e. host-based as oppose to random-based analysis).

Due to the large size of the host genome, the analysis of under-represented sequences in the hosts was performed differently than in the viruses. Instead, the hosts were analysed relative to their corresponding viruses. Recall that a host can be infected by several viruses. Specifically, for each pair of a host and a corresponding virus (i.e. a virus that infects that host), we randomly sampled the host coding sequences with a sample size equals the total size of the virus coding sequences. Twenty host samples were used for each host–virus pair. Each sample was compared with 1,000 corresponding randomization variants generated by each of the random models. Thus, twenty sets of under-represented sequences were identified in the host, for each reading frame, given a corresponding virus. A sequence that is under-represented in at least ten of the twenty samples, per reading frame, is then considered as under-represented in the host, given the corresponding virus. This is referred to as the *sampled majority under-represented set* of the host given a corresponding virus (see Supplementary document, Section 1.4.2). The final set of sequences that are under-represented in the host was defined by the intersection over all the corresponding viruses. See more details in Supplementary document, Section 1.4.

2.4 Preparation of synthetic attenuated ZIKV vaccine based on under-represented oligos

The genome of a Thai-strain ZIKV from an infectious-clone plasmid[35] was evaluated to uncover under-represented sequences (see Supplementary document, Section 1.8 for more details on the ZIKV strain). First, the two randomized models (dinucleotides and synonymous codons) were used on the ZIKV coding sequence to identify short sequences that are under-represented. Next, oligos of five nucleotides (5-mers) that were identified by both models and showed significant *P*-values were selected and ranked according to their significance level (see the list of oligos detected in Supplementary document, Section 2.7). Following, the sequence of the Thai strain ZIKV NS5 protein was systematically scanned at the nucleotide level (according to the significance in the relevant frame) to identify locations that can be modified with each 5-mer, but without affecting the amino acid sequence of the protein (Fig. 2). Specifically, we were able to identify and introduce 29 synonymous codon changes in the first reading frame, and 70 synonymous codon changes in the second reading frame.

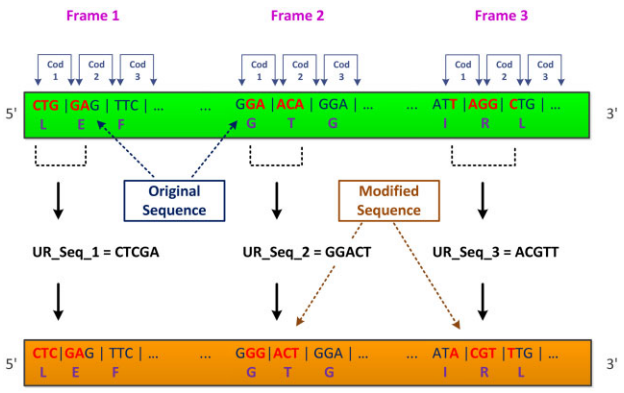

Figure 2. A general scheme of engineering a synthetic sequence. Specifically, in the case of the synthetic ZIKV UR99 sequence, we introduced different under-represented 5-mer oligo in the first two reading frames (identified using both randomization models), replacing the original nucleotide sequence while verifying that the protein AA sequence remains unchanged.

The modified NS5 sequence (hereafter named UR99) was later synthesized as plasmid DNA, amplified by PCR, and used to build ZIKV-UR99 strain by Gibson assembly.[36] The first-passage stock virus was produced using Vero cells.

Synthetic strain preparation: The infectious-clone plasmid of the Thai-strain ZIKV was constructed from PCR products of viral cDNA. The transfection of the plasmid into mammalian cells generated infectious virus with replication kinetics similar to those of the original virus. The sequence of the infectious-clone plasmid was indeed verified.[35] The viral sequence from this infectious-clone plasmid was evaluated to uncover under-represented sequences as discussed above.

Cell lines: BHK21 with rtTA3 was used to generate virus from assembled DNA.[37] The supernatant from the transfected BHK21 was then used to infect Vero cells to prepare the virus stock for subsequent experiments. Replication kinetics of the wild-type (WT) virus and the UR99 virus were characterized in Vero cells with MOI = 0.01. The infectious titre was quantitated with Vero cells using immunostaining against E protein by 4G2 monoclonal antibodies.

Animals: The 45 male and female AG129 mice produced by an in-house colony were used. Groups of animals of both genders were randomly assigned to experimental groups and individually marked with ear tags. Animals were challenged with Malaysian ZIKV, ZIKV WT synthetic, UR99, or vehicle. Serum was collected from all mice 14 dpi for assessment of neutralizing antibodies (neutAbs) via PRNT assay. Mice were monitored for mortality and disease signs daily. Individual weights were recorded daily throughout the course of the study.

Virus: WT ZIKV (Malaysian strain, P6-740) was prepared by two passages in Vero cells. A challenge dose of ~100 CCID50 was administered via s.c. injection in a volume of 0.1 ml. The virus was generated from the same infectious-clone plasmid as the designed variants.

Quantification of neutAb: neutAb was quantified using a 50% plaque reduction neutralization titre (PRNT50) assay. Serum samples were heat inactivated at 56°C for 30 min in a water bath. One half serial dilution, starting at a 1/10 dilution of test sera was made. Dilutions were then mixed 1:1 with an appropriate titre of ZIKV in MEM containing 2% fetal bovine serum (FBS) and incubated at 4°C overnight. The virus-serum mixture was then added to individual wells of a 12-well tissue culture plate with Vero76 cells (4e5 cells/well). Viral adsorption proceeded for 1 h at 37°C and 5% CO_2, followed by addition of 1.7% (4,000 cps) methylcellulose overlay medium containing 10% FBS to each well. Plates were incubated for 4 days, and then stained with crystal violet [with 1% (wt/vol) crystal violet in 10% (vol/vol) ethanol] for 20 min. The reciprocal of the dilution of test serum that resulted in >50% reduction in average plaques from virus control was recorded as the PRNT50 value.

3 Results and discussion

3.1 Results

3.1.1 Overview of the study

To identify short under-represented nucleotide sequences, we compared the number of appearances of each 3, 4, and 5 nucleotides sequences in each reading frame of the original genome with many corresponding randomization variants. Our randomization models preserve the basic canonical features of the coding sequences, i.e. amino-acids composition, codon usage bias, and dinucleotide distribution (see Section 2.3). Thus, an under-represented sequence cannot be explained by these canonical features and may be selected against by other evolutionary forces.

To estimate the false discovery rate, we performed two separate evaluations. First, we generated 10,000 randomizations (instead of 1,000) for few randomly selected viruses and verified that under-represented sequences that were detected using 1,000 randomizations were also detected using 10,000 randomizations. In the second evaluation, we performed identifications of under-represented sequences in random variants of the viruses (rather than in the original genome). Specifically, a random variant of each virus was randomly selected, and the *P*-value was evaluated relative to this (random) variant (see Supplementary document, Section 1.4.1.1). Comparing the number of under-represented sequences identified in the original viruses and the randomized variants of the viruses yields an estimation of a false discovery rate of 1.38% (for $m = 3$), 1.39% (for $m = 4$), and 1.43% (for $m = 5$).

The under-represented sequences identified were further processed by analysing different virus and host groups. Specifically, we analysed under-represented sequences for each virus group, for each host domain, for all viruses that corresponds to the same host, and for different combinations of host domains and virus groups (see Supplementary document, Section 1.5). A complete list of the most abundant under-represented sequences among the different virus groups is available in Supplementary Table S2.

In addition, we refined our analysis of under-represented sequences in the viruses by analysing different protein groups. We classified all viral genes into five mutually exclusive functional groups [surface, structural, enzymatic, unknown (unclassified genes), and other (hypothetical genes)] and showed that the selection against short nucleotide sequences depends on the viral protein function.

Finally, we performed a test study using ZIKV, where we engineered under-represented sequences into the genome of an Asian ZIKV and studied their effect both *in vitro* and *in vivo*.

3.1.2 Under-represented sequences appear in many viruses types

Figure 3A and B depicts the average number of under-represented sequences of size $m = 3$, 4, and 5 nucleotides, identified in few subsets of viruses in both the original and random variants of the virus. See Supplementary document, Section 1.5 for details about the different subsets, and Supplementary document, Section 1.4.1.1 for generating random variants of viruses. As shown in the figures, the average number does indeed increase with the sequence size. Also, many under-represented sequences are found in dsDNA viruses that infect bacteria and vertebrate hosts. The average number of under-represented sequences found in the random variants of the viruses is between 1 and 2% of the average number found in the original genome, suggesting a false discovery rate <2%.

Since the genome of dsDNA viruses tend to be on average larger than the genome of RNA viruses, we aimed at evaluating if the larger number of under-represented sequences identified can be simply attributed to a better statistical signal due to the larger nucleotide size of these viruses. A sampling analysis that we performed (see Supplementary document, Section 2.8) suggests that the number of under-represented sequences identified in dsDNA viruses matches their genomic size, when compared with RNA viruses.

A complete list of under-represented sequences of sizes $m = 3$, 4, and 5 nucleotides in all viruses in the database is available in Supplementary Table S3 (random-based) and in Supplementary Table S4 (host-based).

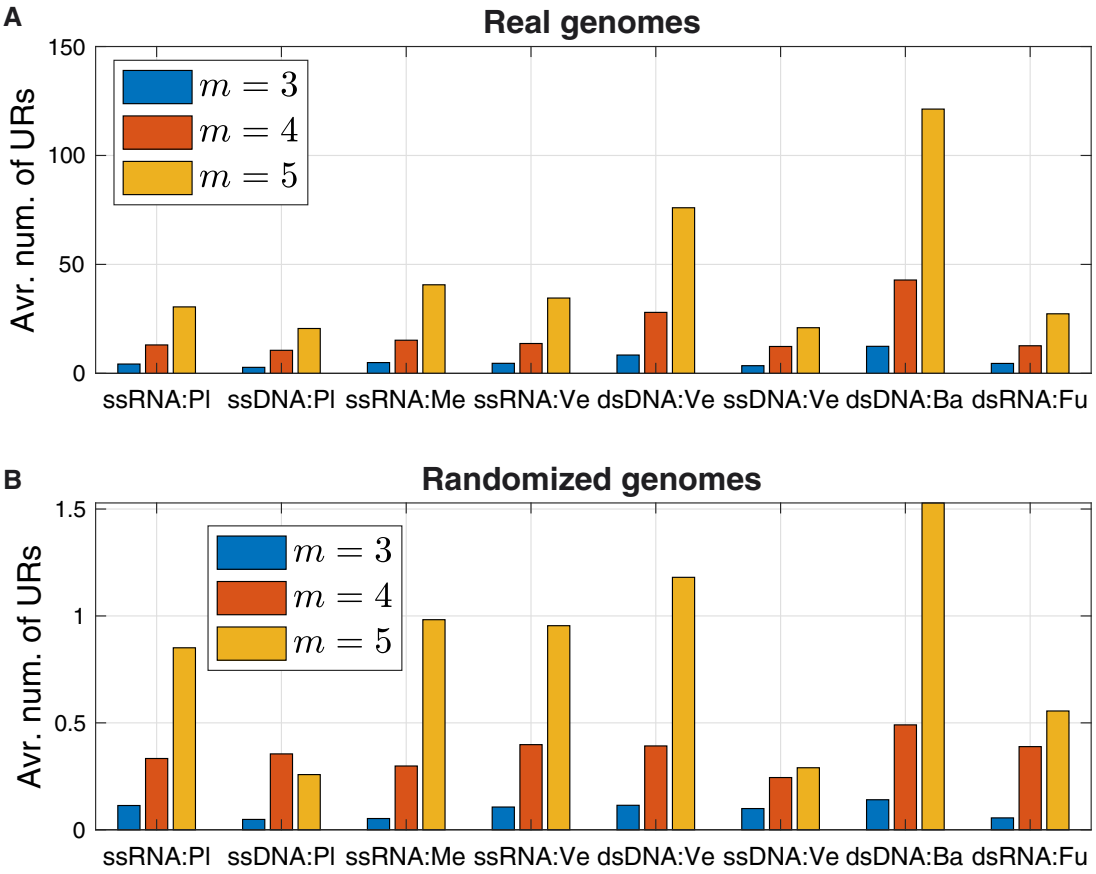

Figure 3. Average number of under-represented sequences of size $m = 3$, 4, and 5 nucleotides. (A) The average number in the original viral genome among different subsets of viruses. (B) The average number in the random viral genome (i.e. in a random variant of the virus) among different subsets of viruses (see Supplementary document, Section 1.4.1.1). The virus's subsets are denoted by a pair V: H, indicating all viruses of type V that infect hosts of domain H (H defines the first two letters of the host domain). For example, ssRNA: Pl denotes all ssRNA viruses that infect hosts of domain plants.

3.1.3 Evidence of universal selection against short homooligonucleotide mainly within the viral coding regions

Our analysis suggests that among the most abundant common under-represented nucleotide sequences (i.e. sequences that are under-represented in all three reading frames) are homooligonucleotide repeats, specifically in viruses. These are sequences of the form XX.X, where all X contain the same nucleotide. Figure 4A depicts the most abundant common under-represented sequences in the five host domains (left figure) and in the five main virus groups (right figure).

Note that among these, specifically in viruses, are sequences containing the same nucleotide repeated $m = 3$, 4, or 5 times (i.e. sequences that correspond to the same colour repeating m times in the figure). A finer resolution of these common under-represented sequences is provided in Fig. 4B, where we depict these sequences separately for different subsets of hosts (left figure) and subsets of viruses (right figure). See Supplementary document, Section 1.5 for more details of the different subsets.

Table 1 lists the six most abundant common under-represented nucleotide sequences of size $m = 3$, 4, and 5 in dsDNA viruses. All homooligonucleotide sequences (shown in red coloured text) are among these most abundant sequences.

One possible reason for this general selection against homooligonucleotide (in all three reading frames) in both viruses and hosts is to reduce erroneous frame shifts as ribosomes traverse the mRNA while decoding it codon by codon. A sequence containing a repetition of the same nucleotide in the coding sequence may cause the ribosome to miss the codon boundary, resulting in a frame shift and thus a non-functional and most likely deleterious protein.[2,38] This must be recognized and degraded by energy-consuming intracellular proteolytic mechanisms. Since translation is the most energetically consuming process in the cell, it is believed that transcripts undergo selection to minimize this energy cost.[39–43] Selection against sequences of repetitive nucleotides reduces faulty translation, thus minimizing the overall translation cost.

It is possible that this selection against homooligonucleotide repeat is indeed more pronounced in viruses than in hosts since viruses are under much stronger evolutionary selection as they have a larger effective population size and thus a stronger effect of these types of mutations on their fitness. Another possible reason may be related to different host immune evasion mechanisms used by viruses (see Section 3.2).

We also evaluated the sequence overlap between common under-represented sequences in viruses and transcription factor binding sites and again found a general selection against homooligonucleotide repeats. These are reported in Supplementary document, Section 2.4.

3.1.4 Evidence of selection against short palindromic sequences within the viral coding regions

A nucleotide sequence is called palindromic if it is identical to its reverse complement. Obviously, palindromic sequences are of even

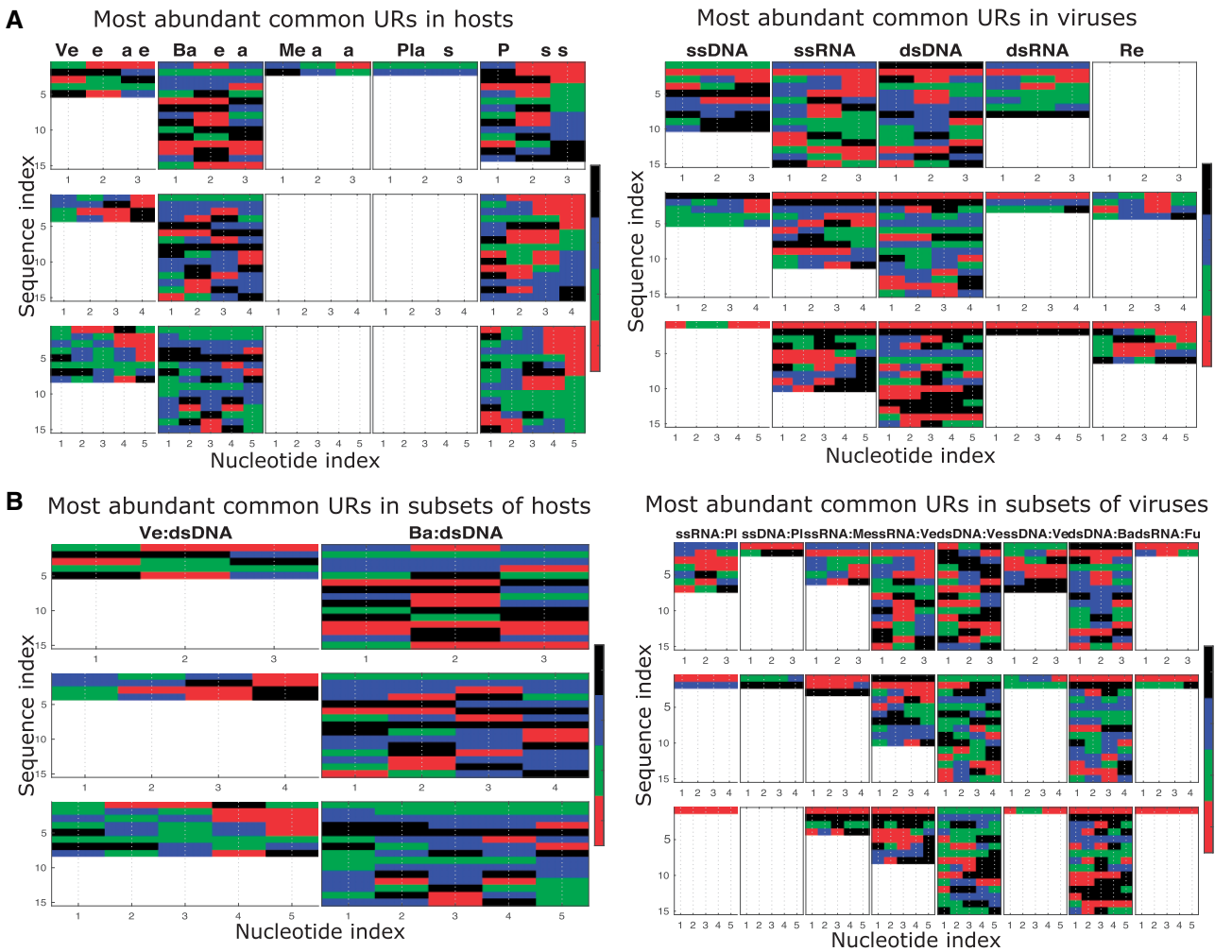

Figure 4. The most abundant common under-represented sequences of size $m=3$ (top panel in each sub-figure), $m=4$ (middle panel in each sub-figure), and $m=5$ (bottom panel in each sub-figure). (A) In five host domains (left) (no common under-represented sequences were found for hosts of the fungi domain) and in the main five virus groups (right). (B) In subsets of hosts (left) and subsets of viruses (right). The host subsets are denoted by the pair H: V, indicating all hosts of domain H that are infected by viruses of type V (H defines the first two letters of the host domain). For example, Ve: dsDNA denotes all hosts of the domain vertebrate that are infected by viruses of type dsDNA. The virus subsets are denoted by the pair V: H, indicating all viruses of type V that infect hosts of domain H. For example, ssRNA: Pl denotes all ssRNA viruses that infect hosts of domain plants. Each row in each panel denotes a nucleotide sequence. A maximum of 15 sequences are shown in each panel ordered top to bottom based on their occurrence frequency (i.e. top sequence appeared most frequently as common under-represented).

Table 1. The six most abundant common under-represented nucleotide sequences of size $m=3$, 4, and 5 in dsDNA viruses (and their abundance in percentage)

$m=3$	$m=4$	$m=5$
TTT (25.5%)	AAAA (25.2%)	AAAAA (29.5%)
AAA (18.2%)	TTTT (24.8%)	TTTTT (22.8%)
TAG (13.2%)	GATC (24.2%)	GGATC (12.8%)
CCC (10.0%)	CGCG (11.1%)	GATCT (11.7%)
GAC (8.7%)	GGGG (10.5%)	GGGGG (11.1%)
GGG (8.6%)	CCCC (8.2%)	CCCCC (10.5%)

Red colour indicates homooligonucleotide repeats.

length. Our analysis reveals that 32.5% of all common under-represented sequences of size $m=4$ nucleotides in viruses are palindromes. Excluding homooligonucleotide repeats this becomes ~51%. Note that only 6.25% of all possible sequences of size $m=4$ nucleotides are palindromes. We also evaluated the number of palindromes in random variants of the viruses. These random variants preserve basic transcript features such as amino-acid order and content, codon usage bias and dinucleotide distributions. Only 5.7% of all common under-represented sequences of size $m=4$ in the random variants of the viruses were found to be palindromes. These findings suggest that indeed the coding regions of viruses are selected against short palindrome sequences.

Figure 5A and B depicts the percentage of palindromic sequences of size $m=4$ nucleotides that are common under-represented sequences in subsets of hosts and viruses. It was found that palindromic sequences are selected against only in one subset of hosts: bacterial hosts that are infected by dsDNA viruses. In addition, palindromic sequences were found to be selected against in dsDNA viruses that infect either bacteria (i.e. bacteriophage) or vertebrate hosts.[44,45]

As depicted in Fig. 5A and B, the sequence GATC is the most abundant palindromic common under-represented sequence in bacteriophages. GATC is a recognition site of different

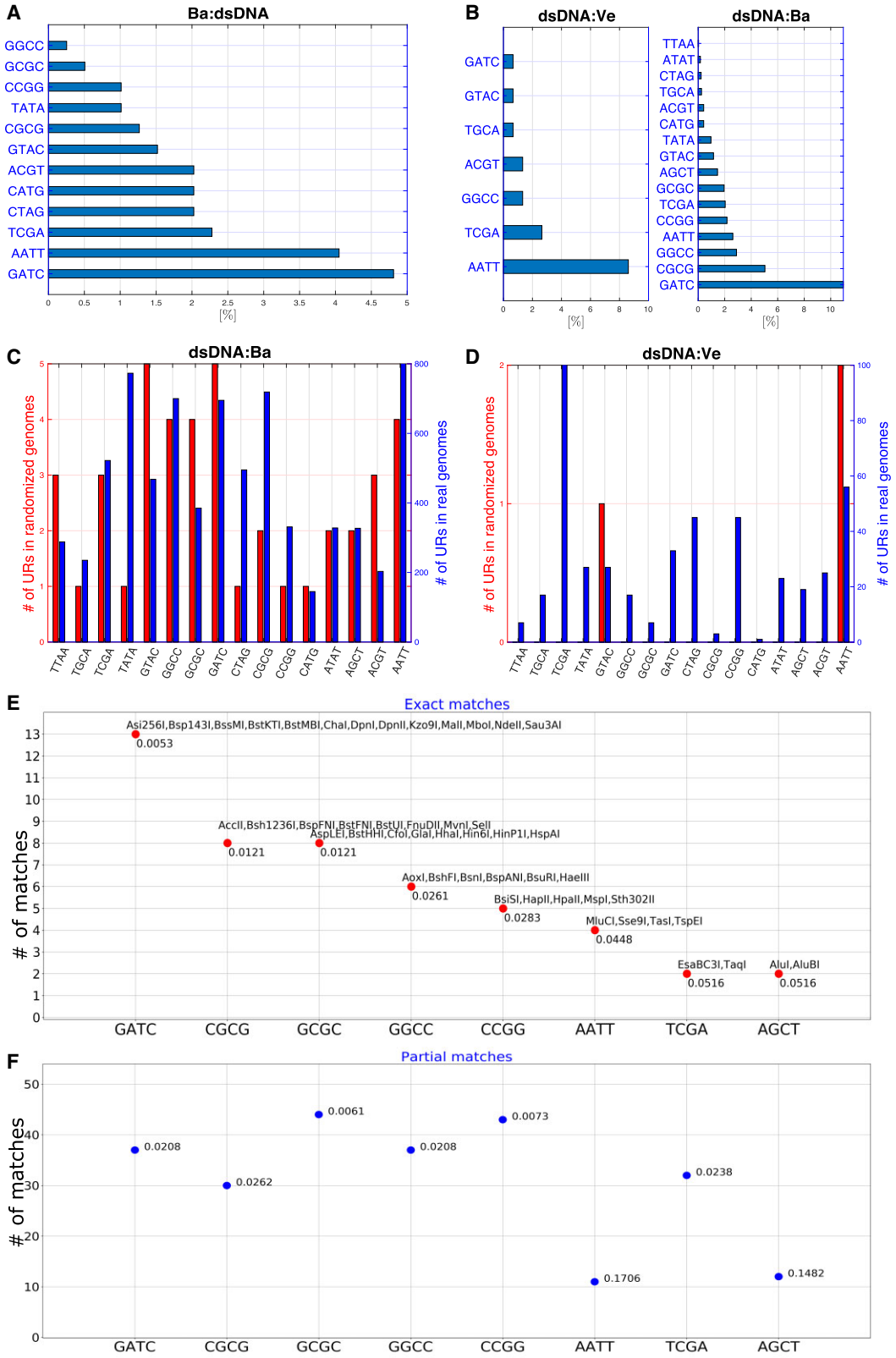

Figure 5. Under-represented palindrome sequences. (A) The percentage of palindromic sequences of size $m=4$ nucleotides that are common under-represented in hosts of domain bacteria that are infected by viruses of type dsDNA. (B) The percentage of palindromic sequences of size $m=4$ nucleotides that are common under-represented in viruses of type dsDNA infecting hosts of domain vertebrate (left) and hosts of domain bacteria (right). (C) The number of occurrences of each palindrome of size $m=4$ as under-represented sequence in viruses of type dsDNA infecting hosts of domain bacteria in the original viral genome (blue) and in the randomized genome (red) of viruses. Note that the scales of the blue and the red bars are extremely different. (D) The number of occurrences of each

restriction-modification systems, as well as solitary methyltransferase Dam. In addition, methyl-directed Type II DpnI enzyme cleaves methylated GATC sequences. Previous work[28] hypothesized that GATC avoidance in bacteria can result from a DNA exchange between strains with different methylation status of GATC site within the process of natural transformation (see also[32]).

Figure 5C and D depicts the total number of occurrences of each palindrome as under-represented sequence in dsDNA viruses that infect bacteria and vertebrate hosts, respectively. In these sub-figures we analysed under-represented sequences regardless of reading frames. Two cases are shown: the case where the real virus genome is used (shown in blue colour), and the case where a randomized variant of the virus genome is used (shown in red colour). Note the scale difference in the *y*-axis between the real and the randomized results. The results in the figures imply that dsDNA viruses undergo selection against short palindrome sequences.

It has been proposed that the principal underlying reason for the apparent avoidance of short palindromes in dsDNA viruses is because they are targets for many restriction-modification systems and possibly for general recombination systems as well.[25,29,31,46,47] Restriction-modification systems protect bacteria and archaea from attacks by bacteriophages and archaeal viruses. A restriction-modification system specifically recognizes short sites in foreign DNA and cleaves it, while such sites in the host DNA are protected by methylation.

To evaluate the hypothesis of palindromes avoidance in viruses due to restriction-modification systems, we downloaded all restriction enzyme patterns from the REBASE[48] database (we used version 811, which contains information for 952 different restriction enzymes) and evaluated the overlap between the common under-represented nucleotide sequences we identified and the restriction sites from REBASE. Figure 5E depicts the number of exact matches between the most abundant common under-represented palindrome sequences of size $m=4$ nucleotides in dsDNA viruses and restriction sites. The figure also depicts the corresponding enzyme name and the *P*-value for each common under-represented sequence. The *P*-value was computed by evaluating the match between common under-represented sequences of random variants of the viruses and the restriction sites. Figure 5F depicts the number of restriction sites that are supersets of the most abundant common under-represented palindrome sequences. *P*-Values were computed as in the case of an exact match.

To show that the correspondence between selection against short palindromic sequences in viruses and restriction sites cannot be explained by basic coding region features such as amino-acid content and order, codon usage bias and dinucleotide distribution, we also evaluated the overlap between restriction sites and common under-represented sequences of random variants of viruses. This is reported in Supplementary document, Section 2.5. A complete list of all common under-represented palindromes of size $m=4$ is provided in Supplementary Table S6.

3.1.5 Large numbers of common under-represented sequences are found in dsDNA viruses infecting vertebrate or bacteria hosts

Figure 6 depicts the number of common under-represented nucleotide sequences identified in different subsets of hosts and viruses. Common under-represented sequences were only identified in two subsets of hosts. On the other hand, common under-represented sequences were identified in all eight subsets of viruses.

Our analysis reveals that dsDNA viruses infecting bacteria and vertebrate hosts have the largest number of common under-represented sequences among the different virus subsets. This, as suggested above, seems to be due to the size of dsDNA viruses when compared with ssDNA and RNA viruses. On the other hand, bacteria that are infected by dsDNA viruses have the largest number of common under-represented sequences among the different host subsets. Thus, the stronger selection for under-represented sequences in bacteria may induce stronger selection for under-represented sequences in viruses that utilize this host.

In addition, we evaluated the number of under-represented sequences identified in the real genome of the viruses when compared with the randomized genome of the viruses. This is reported in Supplementary document, Section 2.9. Indeed, many more sequences are identified as under-represented in the real genome of the virus. On average over all viruses and the three sequence sizes, there are ~45 STDs more under-represented sequences in the real genome in comparison to the random genomes, implying that these cannot be explained by basic coding region features, and suggesting possibly new evolutionary forces acting on the viral coding regions.

3.1.6 Many sequences are common under-represented in viruses but not in their related hosts

We analysed the correspondence of the under-represented nucleotide sequences between hosts and their related viruses. Specifically, for each pair of a host and a corresponding virus we identified three different classes of sequences:

A. Sequences that are common under-represented in both the host coding regions and in the corresponding virus coding regions.
B. Sequences that are common under-represented in the corresponding virus coding regions but are **not** common under-represented in the host coding regions.
C. Sequences that are common under-represented in the host coding regions but are **not** common under-represented in the corresponding virus coding regions.

Note that since we analyse each pair of a host and a corresponding virus separately, the set of under-represented sequences in a host above is the sampled majority under-represented set.

For obvious reasons, sequences that are not under-represented in both host and virus coding regions constitute the majority of the sequences and are thus not reported here. A complete list of all under-represented sequences within the three classes above for all hosts and viruses in our database is available in Supplementary Table S5.

In general, an under-represented sequence of m nucleotides may contain sub-sequences that are themselves under-represented. Thus, it may be interesting to identify unique under-represented sequences, i.e. sequences that do not contain any sub-sequences that are under-represented. For each pair of a host and a corresponding virus, a sequence belonging to one of the three classes above is referred to as a

palindrome of size $m=4$ as under-represented sequence in viruses of type dsDNA infecting hosts of domain vertebrate in the original viral genome (blue) and in the randomized genome (red) of viruses. Note that the scales of the blue and the red bars are extremely different. (E) Overlap between common under-represented sequences of size $m=4$ nucleotides in dsDNA viruses and restriction sites downloaded from the REBASE database. Shown are the number of exact matches between the most abundant common under-represented palindromes of size $m=4$ in dsDNA viruses and restriction sites. The corresponding restriction enzyme names and *P*-values are shown as well. (F) The number of restriction sites that are a superset of the most abundant common under-represented palindromes of size $m=4$ nucleotides in dsDNA viruses. Shown also are the corresponding *P*-values.

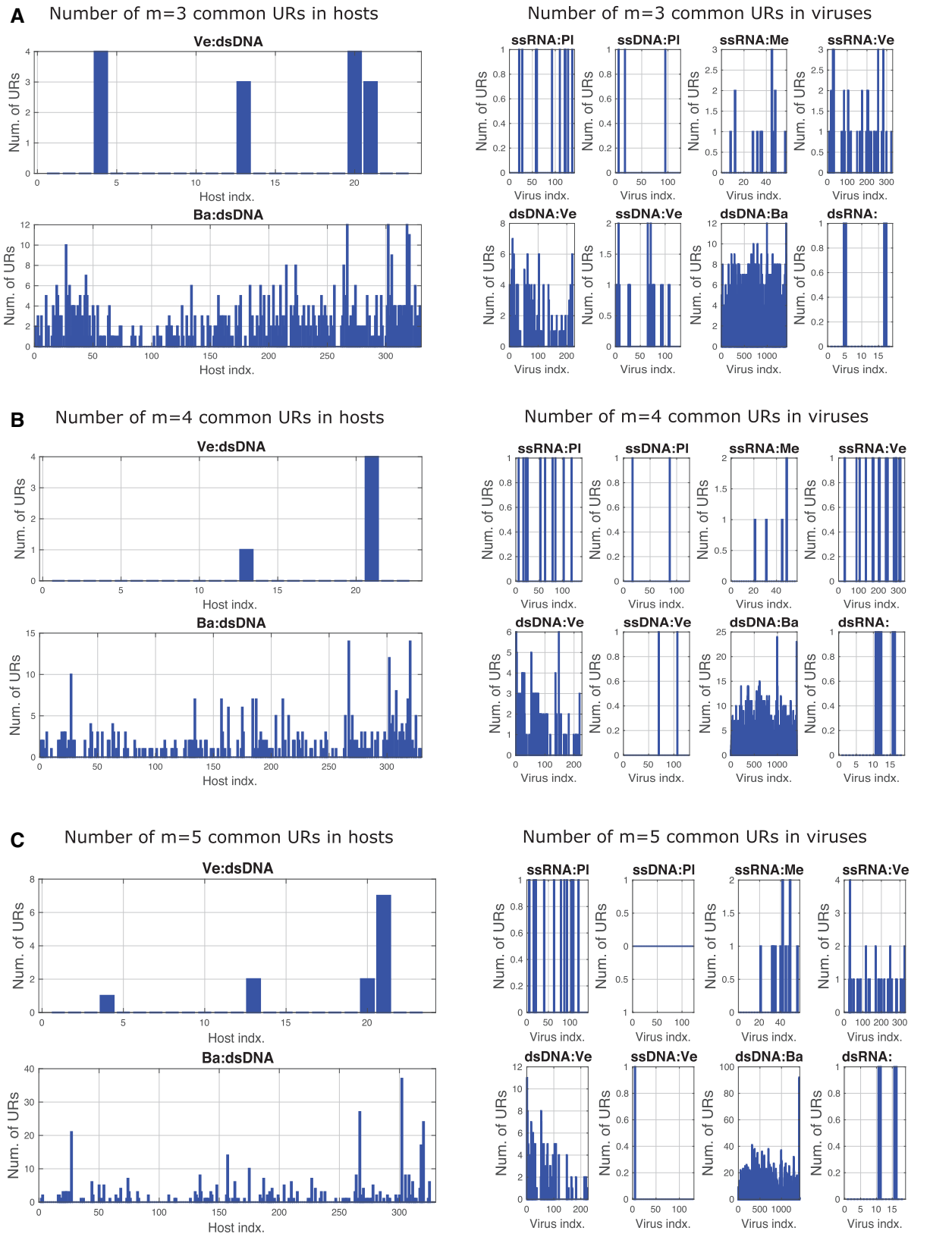

Figure 6. The number of the common under-represented nucleotide sequences in subsets of hosts and in subsets of viruses. A, B, and C correspond to sequences of size $m = 3$, 4, and 5 nucleotides, respectively, where in each panel the left sub-figure corresponds to subsets of hosts and the right sub-figure to subsets of viruses.

unique under-represented sequence if it does not contain any sub-sequence that is under-represented in that class. Specifically, a unique common under-represented sequence of size $m=4$ ($m=5$) nucleotides doesn't contain any sub-sequence of size $m=3$ (of size $m=3$ and of size $m=4$) nucleotides that is common under-represented sequences. A complete list of all unique common under-represented sequences within the three classes above for all hosts and viruses in the database is available in Supplementary Table S7.

The correspondence of the most abundant under-represented sequences between viruses and their related hosts is depicted in Fig. 7 for different host and virus subsets. Each panel depicts both the most abundant common under-represented sequences (left) and the most abundant unique common under-represented sequences (right), where the panel names correspond to the class names. Our first observation is that many under-represented sequences are indeed unique. For example, comparing the cases of $m=4$ and $m=5$ of class A (left sub-figure middle and bottom rows, respectively) with the corresponding unique set (right sub-figure top and bottom rows, respectively) reveals that the majority of the most abundant sequences is unique. Second, homooligonucleotide repeats are among the most abundant sequences in all three classes. In addition, more sequences were identified in class B over the different subsets than in the other two classes. For example, Table 2 lists the most abundant unique sequence of classes B and C in all the different subsets of hosts and viruses. As shown in the table, unique sequences were identified in all subsets in class B, as oppose to class C.

3.1.7 Selection against under-represented sequences in viruses depends on the protein function

The viral genome encodes different types of proteins that are necessary for the life cycle of viruses in their respective hosts. These, in general, include surface proteins that interact with the host receptors and enable attachment and entry to the host cell, structural proteins that serve as the building blocks of the virus, and replicating enzymes, such as RNA and DNA polymerase, that are required for the replication of the virus. In addition, many other proteins, some of which are uncharacterized, are diversely involved in different regulatory and accessory functions.

Here, our aim is to refine the analysis of under-represented sequences in viruses by analysing, separately, different protein groups. To that end, and similarly to,[21] we classified all viral genes into five mutually exclusive functional groups (functional sets): surface, structural, enzymatic, unknown (unclassified genes), and other (hypothetical genes). Specifically, for each virus in the database, we divided its genome into the five gene sets defined above. Each gene set contains all the virus genes of the same functional group. For example, the surface gene set of a virus contains all the genes that encode surface proteins in the virus's genome. A set might be empty for a particular virus if no genes of the corresponding functional group exist in that virus. See Supplementary document, Table S2 for a list of the total number of sets and genes of each functional group in the database. The analysis of under-represented sequences was then performed separately in each of the five gene sets for each of the viruses in the database (see more details in Supplementary document, Section 1.6). A complete list of all under-represented sequences in each viral functional group over all viruses in the database is available in Supplementary Table S8.

We first analysed the average number of under-represented sequences identified in each gene set. To control for the difference in the average gene size and the number of genes in each set, we randomly selected 1,500, 1,240, 1,450, 3,300, and 2,210 genes from each of the surface, structural, enzymatic, unknown, and hypothetical functional groups, respectively. This means that the number of identified under-represented sequences is analysed over similar region sizes, and the differences between the different sets cannot be explained by the genes' nucleotide size in each set.

Figure 8A depicts the average number of under-represented sequences (over all three reading frames) identified in each of the gene set over the (randomly selected) subset of genes. Relatively small number of under-represented sequences were identified in surface genes (that participate in the recognition of the host receptors), when compared with the other gene sets. At least twice as many were identified in many of the enzymatic genes. These proteins interact closely with the host cell machinery, are essential for the viral replication cycle, and thus must use mechanisms that guarantee their function.

Figure 8B depicts the most abundant common under-represented sequences within each viral functional group. These differ between the different functional groups; however, homooligonucleotide sequences appear among the most abundant common under-represented sequences in all groups.

3.1.8 Under-represented sequences attenuate ZIKV replication *in vitro* and *in vivo*

We designed an attenuated ZIKV variant based on the under-represented analysis we performed. Such variants may be useful in the future for generating a live-attenuated vaccine.

Specifically, we introduced synonymous mutations to the NS5 nucleotide sequence, which includes under-represented sequences, and named the new variant UR99 (see details in Section 2).

Infection studies in Vero cells demonstrated fractional variant attenuation of the UR99 virus, which was correlative with our model predictions (see foci size in Fig. 9A, right bottom). In addition, infectious virus collected and evaluated from the UR99 variant showed substantial attenuation relative to WT ZIKV (Fig. 9A).

There is evidence that AG129 mice lacking IFN-α/β and IFN-γ (types I and II interferon) receptors can be valuable for evaluating the efficacy of new vaccines and anti-viral treatments for ZIKV.[49,50] Therefore, as these mice are immune compromised, various strains of ZIKV cause lethal infection and disease, and will typically cause morbidity and mortality. Depending on the strain, severe disease is observed between 1 and 2 weeks after virus challenge.[50,51]

Thus, to further test the synthetic vaccine attenuation level *in vivo*, AG129 mice were challenged with attenuated ZIKV preparations as well as synthetic WT ZIKV. These inoculations were done in parallel with the original virus grown in cell culture. Infection with the synthetically attenuated ZIKV strains was lethal in all inoculated AG129 mice. However, the mortality curve of mice infected with UR99 was delayed, when compared with that of WT Malaysia and WT synthetic ZIKV (average of 20.4 days in UR99 *vs*. 15 and 17.5 in WT Malaysia/synthetic ZIKV, respectively; see Fig. 9B). No mortality was observed in unvaccinated controls, and mice vaccinated with vehicle (Fig. 9B).

Weight loss was also observed in all the infected mice (30–40%; see Fig. 9C). Normal control mice experienced general weight gain throughout the experimental period (Fig. 9C). Weight loss corresponded well with mortality, and mice typically lost substantial weight, requiring humane euthanasia.

neutAb is the primary mediator of protection in vaccine studies in this model.[52,53] Therefore, serum samples were taken to determine the presence of neutAb in infected mice. The neutAb titre was evaluated in vaccinated mice 2 weeks after vaccination. Mice vaccinated with synthetic WT or UR99 had significantly ($P<0.0001$) elevated

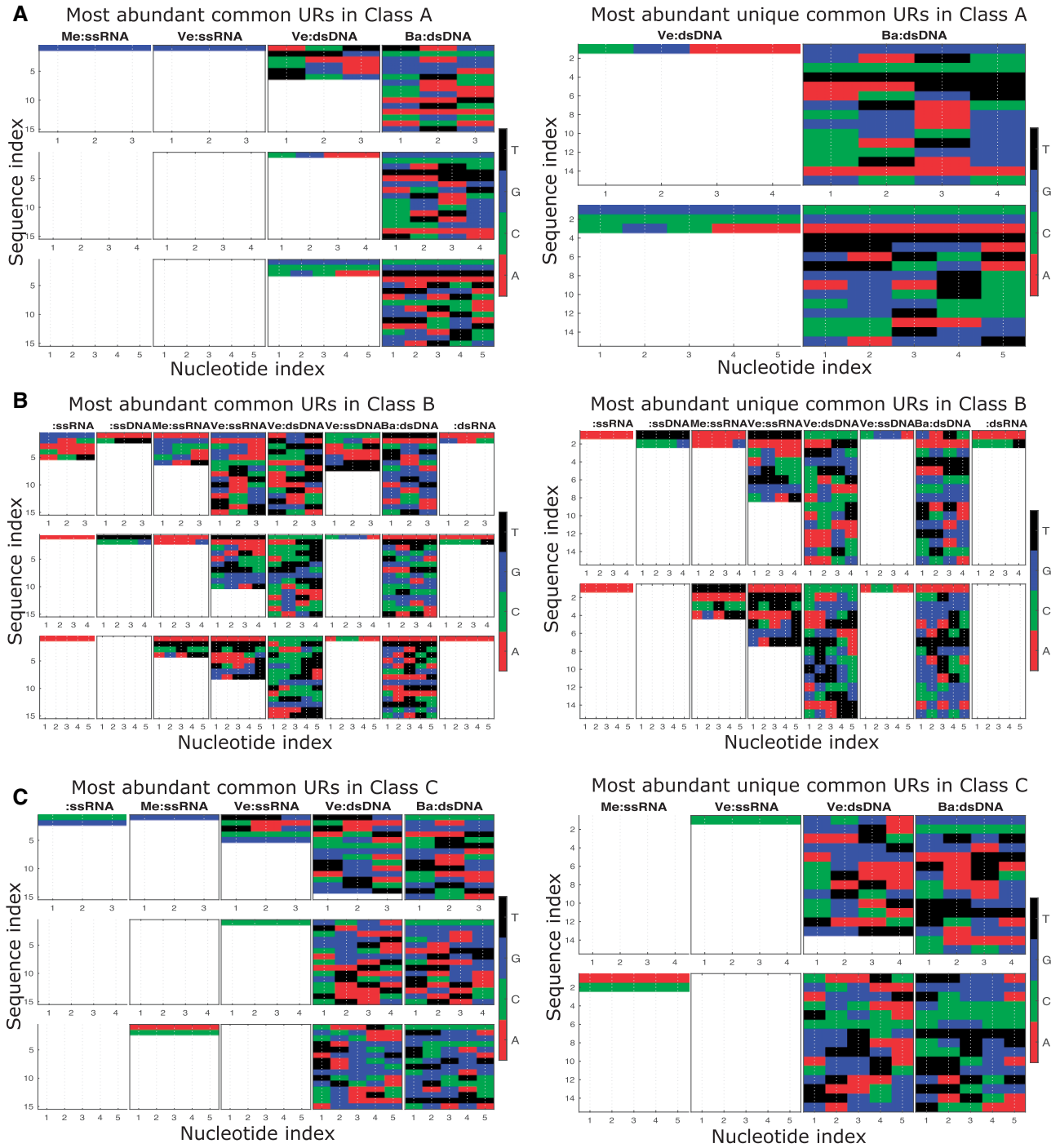

Figure 7. The most abundant common under-represented nucleotide sequences that are shared between hosts and their corresponding viruses in different subsets of hosts and viruses. (A) Class A sequences (left) of size $m=3$ (top panel), $m=4$ (middle panel), and $m=5$ (bottom panel), and unique class A sequences (right) of size $m=4$ (top panel) and $m=5$ (bottom panel). (B) Class B sequences (left) of size $m=3$ (top panel), $m=4$ (middle panel), and $m=5$ (bottom panel), and unique class B sequences (right) of size $m=4$ (top panel) and $m=5$ (bottom panel). (C) Class C sequences (left) of size $m=3$ (top panel), $m=4$ (middle panel), and $m=5$ (bottom panel), and unique class C sequences (right) of size $m=4$ (top panel) and $m=5$ (bottom panel). Each row in each panel denotes a nucleotide sequence. A maximum of 15 sequences are shown in each panel ordered top to bottom based on their occurrence frequency (i.e. top sequence appeared most frequently as common under-represented).

neutAb titres when compared with vehicle controls (see Fig. 9D). As expected, no neutAb was detected in mice vaccinated with vehicle or in normal control groups (see Fig. 9D).

The virulence levels of UR99 were somewhat lower than the levels of the Malaysian and Synthetic WT strains, thus demonstrating that under-represented sequences can be potentially used in the design of live attenuated ZIKV strains. Accordingly, additional attenuation of this variant (e.g. by introducing similar changes to other ZIKV proteins) may further decreases the lethality of the mice infected by it. Since AG129 mice are very susceptible to ZIKV infection,[49–51] this

Table 2. The most abundant sequence that is unique common under-represented (of size $m=4$ and $m=5$) in viruses but not in the corresponding hosts (top row), and in hosts but not in the corresponding viruses (bottom row)

Pl-ssRNA	Pl-ssDNA	Me-ssRNA	Ve-ssRNA	Ve-dsDNA	Ve-ssDNA	Ba-dsDNA	Fu-dsRNA
AAAA (6.7)	TTTT (1.6)	AAAA (4.2)	TTTT (1.1)	CCCC (11.1)	CGGA (0.8)	GATC (22.2)	AAAA (1.1)
AAAAA (7.5)		TTTTT (8.3)	AAAAA (5.3)	CCCCC (15.3)	ACCAA (0.8)	AAAAA (8.2)	
X	X	AAAAA (2.1)	CCCC (0.9)	GCGA (12.1)	X	GGGG (16)	X
				CAATC (3.1)		TTGGA (9.6)	

The numbers in parenthesis indicate the frequency of occurrences in percentage. X indicates that no corresponding sequence was identified.

mouse model might be too stringent to test these live attenuated vaccine candidates, as human infection is generally sub-clinical after natural ZIKV infection, hence the attenuated strain might be effective in an immunocompetent model.

3.1.9 Average number of under-represented sequences in viruses and hosts

We compared the average number of under-represented sequences identified in each pair of a virus and its corresponding host. See Supplementary document, Section 2.10 for more details. We found that in ∼75% of the cases the average number was larger in the hosts. We believe that this is due to the fact that the viral genome is usually populated with many overlapping codes and genes, when compared with cellular organisms.[54–56] This introduces many constraints along the viral genome, which can decrease the number of under-represented sequences in the virus. For example, a sub-optimal codon within the host coding region may be synonymously replaced by evolution without affecting the host fitness. However, due to overlapping codes, replacing a sub-optimal codon within the viral coding region may affect multiple proteins and genes, and thus be deleterious to the virus.

3.2 Discussion

In this study, we analyse sequences of three, four, and five nucleotides long that are under-represented in the coding regions of viruses of all types and in their corresponding host coding regions. This study is based on a novel statistical evaluation that controls for classical coding region features, which is performed separately in each of the three reading frames. We provide various novel discoveries that may shed light on the evolution of viral DNA sequences and on the virus co-evolution with its respective hosts. It is important to emphasize that the observed patterns may be related to various variables and their complex interactions, include gene expression optimizations, various mechanisms for escaping the host immune system, and co-evolution with the corresponding hosts. For example, it was reported that suppression of CG dinucleotides in HIV-1 is due to co-evolution with its vertebrate host to avoid the host defence mechanisms.[18]

In general, our analysis reveals that under-represented viral sequences are related to different mechanisms such as restriction modification systems and possibly to alternative or unknown immune escape mechanisms, as these sequences cannot be explained by canonical mechanisms that may suggest, for example, classical viral recognition using antibodies.

We show that homooligonucleotide repeats are the most abundant under-represented sequences in both viruses and hosts. A possible explanation for this avoidance is to reduce an erroneous ribosomal frame shifts and thus reduce faulty translation and consequentially the overall translation cost. However, as this motif is shown to be shared between hosts and viruses, our analysis also indicates that a stronger selection pressure against these sequences exists in viruses. This again can be attributed to escape mechanisms from the host immune system, as the virus nucleotide composition evolves to be similar to the host, and it is certainly possible that an excess avoidance of homooligonucleotide repeats reduces viral recognition by classical host immune mechanisms. There may be other relevant explanations such as interaction with small RNA genes (e.g. miRNAs). It is possible, for example, that these sequences may increase the efficiency of miRNA and mRNA interactions and thus decrease expression levels. This should be studied further.

In addition to homooligonucleotide repeats, we show that palindromes are among the most abundant under-represented sequences in viruses. Specifically, excluding homooligonucleotide repeats, our analysis reveals that 51% of all under-represented sequences of four nucleotides long in viruses are palindromes (where only 6.25% of all possible sequences of that size are palindromes).

Indeed, analysis of palindromes avoidance in viruses was performed previously. It was shown that palindromes are the most under-represented short sequences in a prokaryotic genome.[25–27] For example, it was reported that short palindromic sequences are avoided at a statistically significant level in the genomes of several bacteria.[25] Four and six nucleotides palindromic sequences that are avoided were reported for few viruses and hosts in,[57] and avoidance of palindromes in several dozen phage genomes was reported in[26] These analyses are based on statistical counts of certain sequences in the given DNA and thus do not control for canonical coding region features (codon usage bias, amino acid order and content and dinucleotide distribution) as was done in this study. In addition, our analysis is performed over a larger set of viruses of all types and their corresponding hosts, and at a reading frame resolution. Thus, we believe that the results reported here may be more accurate, and should provide a better understanding of this phenomenon.

One plausible explanation for avoidance of palindromes in viruses is because they are targets for many restriction-modification systems and possibly for general recombination systems as well. We statistically show a high overlap between under-represented palindromes in viruses and restriction enzyme patterns. This overlap cannot be explained by classical coding region features. Restriction of recognition sites has been observed in genomes of prokaryotic organisms.[26,28–30,46] The authors in[29] analysed the avoidance of restriction sites in few bacteriophage, and concentrated on sites containing six nucleotides. Rusinov et al.[46] studied most known recognition sites (both palindromic and asymmetric) in thousands of prokaryotic genomes and found factors that influence their avoidance. It was also shown that the recognition site avoidance correlates with the lifespan of restriction-and-modification systems. Recently, the authors in[31]

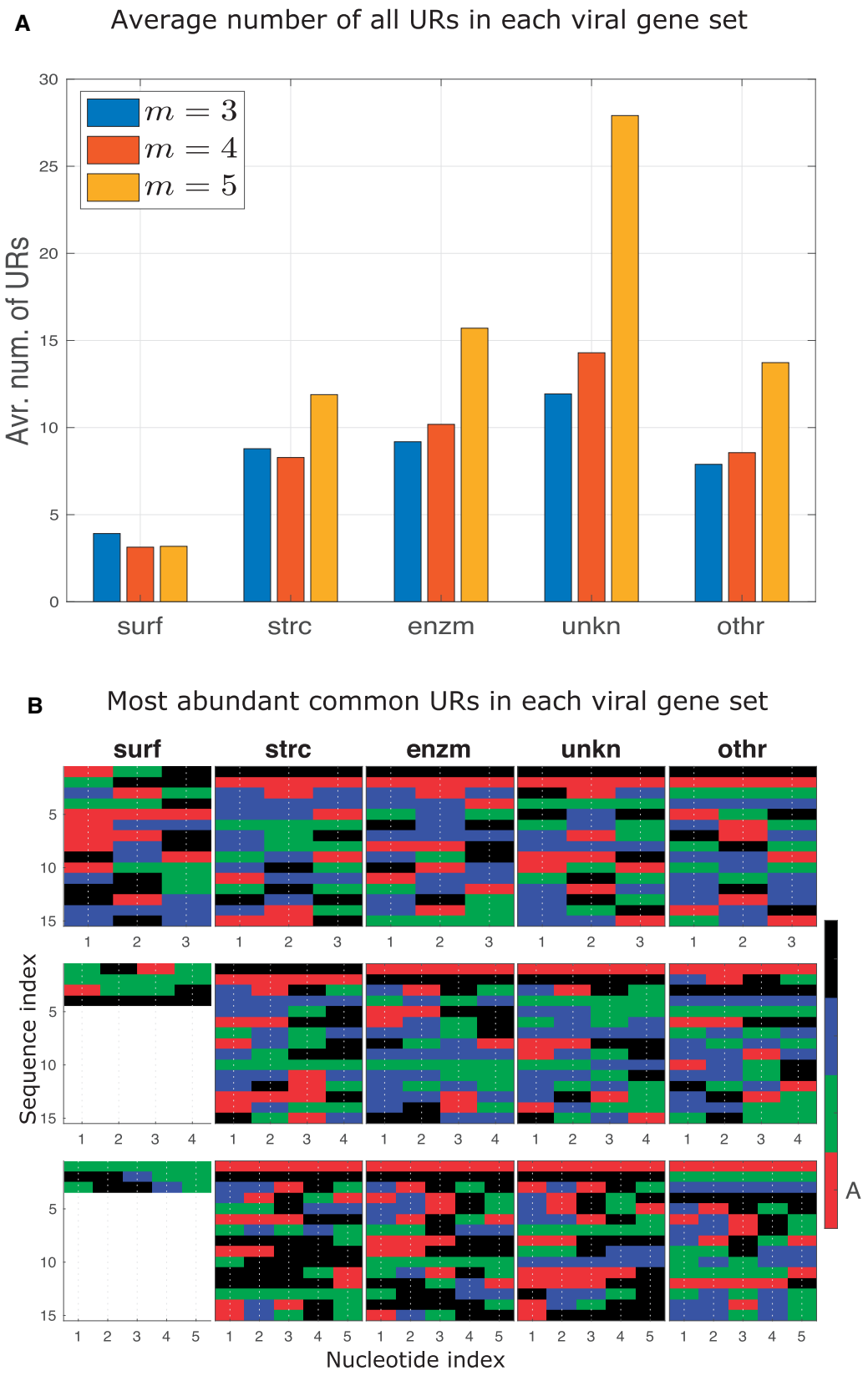

Figure 8. Under-represented sequences within the virus functional gene sets. Here, 'surf' stands for surface, 'strc' for structural, 'enzm' for enzymatic, 'unkn' for unknown (unclassified), and 'othr' for other (hypothetical) functional groups. (A) The average number of under-represented sequences, over all three reading frames, of size $m = 3$, 4, and 5 nucleotides, identified in each viral gene set when analysing (randomly selected) 1,500, 1,240, 1,450, 3,300, and 2,210 genes from each of the surface, structural, enzymatic, unknown, and hypothetical functional groups, respectively. (B) The most abundant common under-represented nucleotide sequences in each of the virus functional group, of size $m = 3$ (upper panel), $m = 4$ (middle panel), and $m = 5$ (lower panel). Each row in each panel denotes a nucleotide sequence. A maximum of 15 sequences are shown in each panel ordered top to bottom based on their occurrence frequency (i.e. top sequence appeared most frequently as common under-represented).

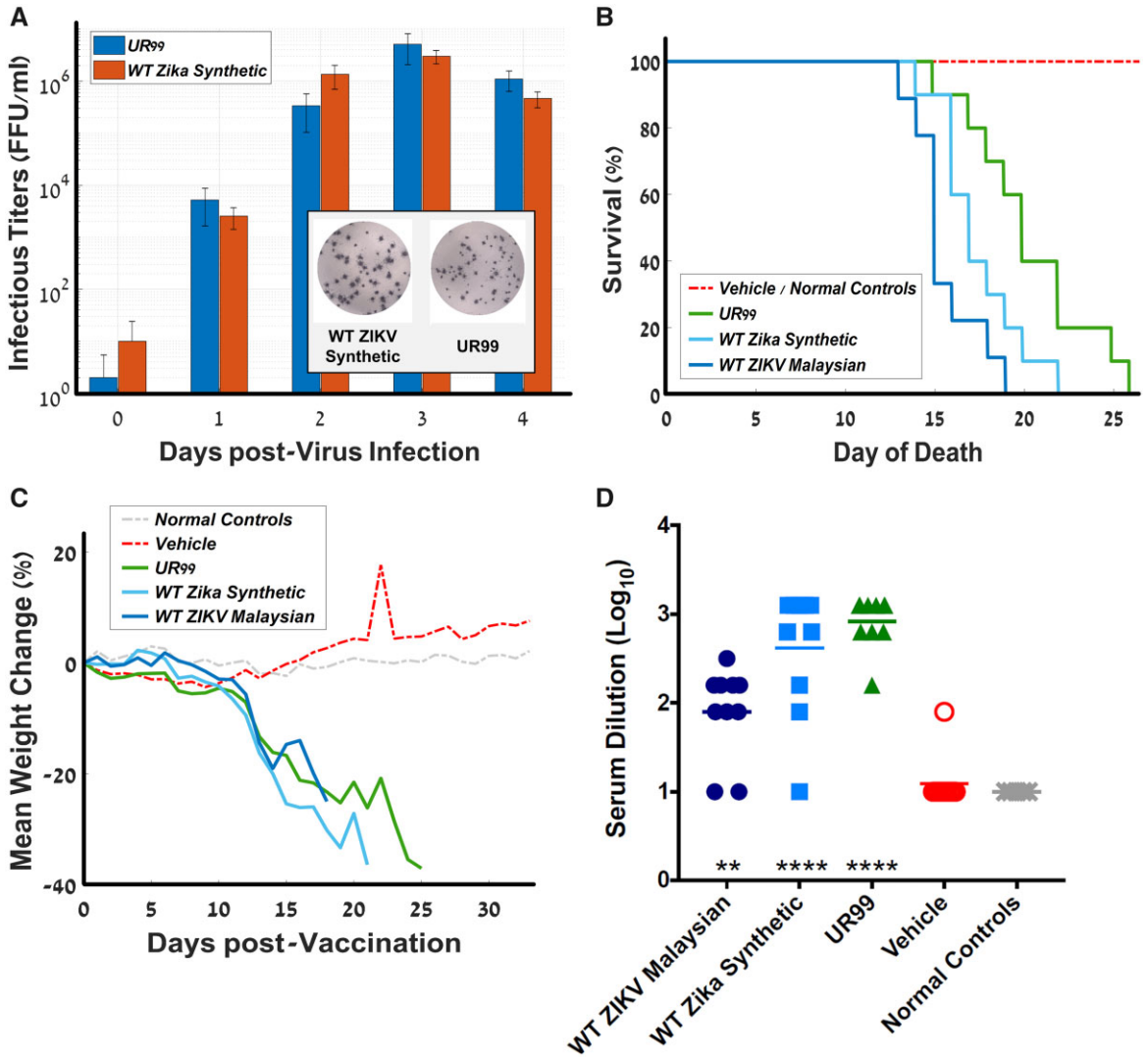

Figure 9. Incorporation of under-represented sequences produced an attenuated ZIKV variant. (A) Foci size and replication kinetics of WT ZIKV and UR99 in Vero cells. The smaller foci size comparison demonstrates variant attenuation of the UR99 (bottom right). Titre analysis shows the UR99 variant attenuation relative to WT ZIKV (borderline significant *P*-value in Day 2: 0.078). (B) Mortality curves of AG129 mice infected with UR99, synthetic WT ZIKV, or Malaysian strain ZIKV. (C) Average weight change, in percentage, of animals infected with WT ZIKV Malaysian, synthetic WT ZIKV, or UR99. (D) PRNT50 titres from serum collected from vaccinated AG129 mice 13 days post vaccination (****$P < 0.0001$ and **$P < 0.01$ when compared with vehicle treatment).

analysed avoidance of recognition sites of restriction-modification systems in the genomes of prokaryotic viruses and found it to be a widespread but not a universal anti-restriction strategy of these viruses. The method used by the authors is based on a compositional bias calculation, which is the ratio of the observed to the expected frequency of a sequence, where the expected frequency is estimated based on the observed frequencies of all sub-sites of a given sequence. The compositional bias measure was originally used in[32] for analysing over- and under-represented sequences in DNA viruses. Since the compositional bias measure doesn't account for a statistical background that preserves know evolutionary forces, we believe that a more accurate and comprehensive procedure of identifying under-represented sequences is the one used here.

In addition, we analyse the distribution of these under-represented sequences among various viral and host groups. We show, for example, that dsDNA viruses infecting bacteria or vertebrate hosts contain a larger set of under-represented sequences than other viral types and that this may be related to their larger genome size. Furthermore, we show that on average the set of sequences that are under-represented in viruses but are not under-represented in their related hosts is the largest set among different host–virus under-represented correspondence.

We also show that the selection against under-represented sequences in viruses depends upon the protein function. For example, larger number of sequences is shown to be under-represented in enzyme genes than in surface genes. Moreover, even larger number of sequences is found to be under-represented in genes with (currently) unknown functionality, prompting further investigation into the nature of these genes. The differences between these groups may also be related to the expression levels of the different proteins. If, for example, surface genes tend to have low expression levels then they may be under weaker selection for features such as under-represented sequences.

Vaccines are a topic of a singular importance in present day biomedical science. However, the discovery of vaccines has so far been

primarily empirical in nature requiring considerable investments of time, efforts, and resources.[58] To overcome the numerous pitfalls attributed to the classical vaccine design strategies, more efficient and robust rational approaches are highly desirable. One direction in designing *in silico* vaccine candidates may be based on exploiting the synonymous information, encoded in the viral genomes and related to gene expression, for attenuating the viral replication cycle while retaining its genotype and structure. The analysis and results reported here may have important implications in vaccine synthesis. Specifically, the outcomes of this study may provide clues and guidance into practical design of efficient and safe viral vaccines via attenuated viral material. Furthermore, it may also prove to be beneficial for other biotechnological objectives related to viral based products such as developing oncolytic viruses and engineering phages to fight bacteria.[59–64] Indeed, we demonstrate, both *in vitro* and *in vivo*, how under-represented sequences can be utilized to obtain an attenuated ZIKV.

The aim of these experiments is an initial proof of concept. Of course, additional experiments with more variants and controls are needed to better understand the effect of these under-represented sequences on the viral growth rate and fitness. For example, it will be helpful to study additional mutants that do not possess under-represented sequences but include other types of mutation. However, it is important to emphasize an interesting and a non-obvious aspect of these experiments. The introduced mutations are silent and thus did not alter the encoded protein. Based on our experience, in many cases silent mutations may not affect the viral fitness, and furthermore, there are cases where they may even improve its growth rate. Also, it is important to emphasize that in these experiments both the wild-type and the mutant variants were generated by the same process and from the same infectious-clone plasmid.

Finally, the randomization models used in this study may not completely preserve the viral RNA secondary structure, and thus the selection for under-represented sequences may be partially due to alterations in secondary structures.

Acknowledgements

We are grateful to the anonymous referees for comments that greatly helped in improving this paper.

Accession numbers

BCWF01000001-BCWF01000044

Funding

The work of Y.Z. was supported by the Israeli Ministry of Science, Technology and Space and by the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University.

Ethical approval

The animal research ethics committee at Utah State University approved this research.

Supplementary data

Supplementary data are available at DNARES online.

References

1. Knipe, D., Howley, P., Griffin, D., . 2007, *Fields Virology*. Lippincott Williams & Wilkins: Philadelphia.
2. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. 2008, *Molecular Biology of the Cell*. Garland Science: New York.
3. Walsh, D., Mathews, M.B. and Mohr, I. 2013, Tinkering with translation: protein synthesis in virus-infected cells, *Cold Spring Harb. Perspect. Biol.*, **5**, a012351.
4. Filée, J., Forterre, P. and Laurent, J. 2003, The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies, *Res. Microbiol.*, **154**, 237–43.
5. Koonin, E.V., Makarova, K.S. and Aravind, L. 2001, Horizontal gene transfer in prokaryotes: quantification and classification, *Annu. Rev. Microbiol.*, **55**, 709–42.
6. Koonin, E.V. and Dolja, V.V. 2006, Evolution of complexity in the viral world: the dawn of a new vision, *Virus Res.*, **117**, 1–4.
7. Moreira, D. and Brochier-Armanet, C. 2008, Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes, *BMC Evol. Biol.*, **8**, 12.
8. Widmer, K., Zhu, Y., Williams, J.V., Griffin, M.R., Edwards, K.M. and Talbot, H.K. 2012, Rates of hospitalizations for respiratory syncytial virus, human metapneumovirus, and influenza virus in older adults, *J. Infect. Dis.*, **206**, 56–62.
9. Kitazato, K., Wang, Y. and Kobayashi, N. 2007, Viral infectious disease and natural products with antiviral activity, *Drug Discov. Ther.*, **1**, 14–22.
10. García-Sastre, A. 1998, Negative-strand RNA viruses: applications to biotechnology, *Trends Biotechnol.*, **16**, 230–5.
11. Singh, P., Gonzalez, M.J. and Manchester, M. 2006, Viruses and their uses in nanotechnology, *Drug Dev. Res.*, **67**, 23–41.
12. Campos, G.S., Bandeira, A.C. and Sardi, S.I. 2015, Zika virus outbreak, Bahia, Brazil, *Emerg. Infect. Dis.*, **21**, 1885–6.
13. Musso, D., Nilles, E. and Cao-Lormeau, V.-M. 2014, Rapid spread of emerging Zika virus in the Pacific area, *Clin. Microbiol. Infec.*, **20**, O595–6.
14. World Health Organization, 2020, *Rolling Updates on Coronavirus Disease (COVID-19)*. https://www.who.int/emergencies/diseases/novel-coronavirus-2019.
15. Holmes, E. and Drummond, A. 2007, The evolutionary genetics of viral emergence. In: *Wildlife and Emerging Zoonotic Diseases: The Biology, Circumstances and Consequences of Cross-Species Transmission*. Springer, pp. 51–66.
16. Bahir, I., Fromer, M., Prat, Y. and Linial, M. 2009, Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences, *Mol. Syst. Biol.*, **5**, 1–14.
17. Greenbaum, B.D., Levine, A.J., Bhanot, G. and Rabadan, R. 2008, Patterns of evolution and host gene mimicry in influenza and other RNA viruses, *PLoS Pathog.*, **4**, e1000079.
18. Takata, M.A., Gonçalves-Carneiro, D., Zang, T.M., . 2017, CG dinucleotide suppression enables antiviral defence targeting non-self RNA, *Nature*, **550**, 124–7.
19. Lobo, F.P., Mota, B.E., Pena, S.D., . 2009, Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts, *PLoS One.*, **4**, e6282.
20. Goz, E. and Tuller, T. 2016, Evidence of a direct evolutionary selection for strong folding and mutational robustness within HIV coding regions, *J. Comput. Biol.*, **23**, 641–50.
21. Goz, E., Zafrir, Z. and Tuller, T. 2018, Universal evolutionary selection for high dimensional silent patterns of information hidden in the redundancy of viral genetic code, *Bioinformatics*, **34**, 3241–8.
22. Schbath, S., Prum, B. and de Turckheim, E. 1995, Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences, *J. Comput. Biol.*, **2**, 417–37.
23. Rusinov, I., Ershova, A., Karyagina, A., Spirin, S. and Alexeevski, A. 2018, Comparison of methods of detection of exceptional sequences in prokaryotic genomes, *Biochemistry (Moscow)*, **83**, 129–39.

24. Almirantis, Y., Charalampopoulos, P., Gao, J., . 2017, On avoided words, absent words, and their application to biological sequence analysis, *Algor. Mol. Biol.*, **12**, 5.
25. Gelfand, M.S. and Koonin, E.V. 1997, Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes, *Nucleic Acids Research.*, **25**, 2430–9.
26. Rocha, E.P., Danchin, A. and Viari, A. 2001, Evolutionary role of restriction/modification systems as revealed by comparative genome analysis, *Genome Res.*, **11**, 946–58.
27. Karlin, S. and Cardon, L.R. 1994, Computational DNA sequence analysis, *Annu. Rev. Microbiol.*, **48**, 619–54.
28. Ershova, A., Rusinov, I., Vasiliev, M., Spirin, S. and Karyagina, A. 2016, Restriction-modification systems interplay causes avoidance of GATC site in prokaryotic genomes, *J. Bioinform. Comput. Biol.*, **14**, 1641003.
29. Sharp, P.M. 1986, Molecular evolution of bacteriophages: evidence of selection against the recognition sites of host restriction enzymes, *Mol. Biol. Evol.*, **3**, 75–83.
30. Meisel, A., Schroeder, C., Kupper, D., Krüger, D.H. and Reuter, M. 1995, The significance of distance and orientation of restriction endonuclease recognition sites in viral DNA genomes, *FEMS Microbiol. Rev.*, **17**, 177–84.
31. Rusinov, I., Ershova, A., Karyagina, A., Spirin, S. and Alexeevski, A. 2018, Avoidance of recognition sites of restriction-modification systems is a widespread but not universal anti-restriction strategy of prokaryotic viruses, *BMC Genomics*, **19**, 885.
32. Burge, C., Campbell, A.M. and Karlin, S. 1992, Over-and under-representation of short oligonucleotides in DNA sequences, *Proc. Natl. Acad. Sci. USA*, **89**, 1358–62.
33. Tuller, T., Chor, B. and Nelson, N. 2007, Forbidden penta-peptides, *Protein Sci.*, **16**, 2251–9.
34. Mihara, T., Nishimura, Y., Shimizu, Y., . 2016, Linking virus genomes with host taxonomy, *Viruses*, **8**, 66.
35. Bunpote Siridechadilok, A.S.N. and Jupatanakul, J.P. 2019, Infectious clone plasmid of a Thai-strain Zika virus and its fluorescent reporter system for high-throughput assay and vaccine development.
36. Siridechadilok, B., Gomutsukhavadee, M., Sawaengpol, T., . 2013, A simplified positive-sense-RNA virus construction approach that enhances analysis throughput, *J. Virol.*, **87**, 12667–74.
37. Suphatrakul, A., Duangchinda, T., Jupatanakul, N., . 2018, Multi-color fluorescent reporter dengue viruses with improved stability for analysis of a multi-virus infection, *PLoS One*, **13**, e0194399.
38. Atkins, J.F., Loughran, G., Bhatt, P.R., Firth, A.E. and Baranov, P.V. 2016, Ribosomal frameshifting and transcriptional slippage: from genetic steganography and cryptography to adventitious use, *Nucleic Acids Res.*, **44**, 7007–78.
39. Kurland, C.G. 1992, Translational accuracy and the fitness of bacteria, *Annu. Rev. Genet.*, **26**, 29–50.
40. Subramaniam, A.R., Zid, B.M. and O'Shea, E.K. 2014, An integrated approach reveals regulatory controls on bacterial translation elongation, *Cell*, **159**, 1200–11.
41. Hooper, S.D. and Berg, O.G. 2000, Gradients in nucleotide and codon usage along Escherichia coli genes, *Nucleic Acids Res.*, **28**, 3517–23.
42. Zafrir, Z., Zur, H. and Tuller, T. 2016, Selection for reduced translation costs at the intronic 5' end in fungi, *DNA Res.*, **23**, 377–94.
43. Tuller, T. and Zur, H. 2015, Multiple roles of the coding sequence 5' end in gene expression regulation, *Nucleic Acids Res.*, **43**, 13–28.
44. Volff, J.-N., Körting, C., Froschauer, A., Sweeney, K. and Schartl, M. 2001, Non-LTR retrotransposons encoding a restriction enzyme-like endonuclease in vertebrates, *J. Mol. Evol.*, **52**, 351–60.
45. Herrera, S., Reyes-Herrera, P.H. and Shank, T.M. 2015, Predicting RAD-seq marker numbers across the eukaryotic tree of life, Genome Biol. Evol., **7**, 3207–25.
46. Rusinov, I., Ershova, A., Karyagina, A., Spirin, S. and Alexeevski, A. 2015, Lifespan of restriction-modification systems critically affects avoidance of their recognition sites in host genomes, *BMC Genomics*, **16**, 1084.
47. Gowers, D.M., Bellamy, S.R. and Halford, S.E. 2004, One recognition sequence, seven restriction enzymes, five reaction mechanisms, *Nucleic Acids Res.*, **32**, 3469–79.
48. Roberts, R.J., Vincze, T., Posfai, J. and Macelis, D. 2015, REBASE—a database for DNA restriction and modification: enzymes, genes and genomes, *Nucleic Acids Res.*, **43**, D298–9.
49. Zompi, S., Santich, B.H., Beatty, P.R. and Harris, E. 2012, Protection from secondary dengue virus infection in a mouse model reveals the role of serotype cross-reactive B and T cells, *J. Immunol.*, **188**, 404–16.
50. Aliota, M.T., Caine, E.A., Walker, E.C., Larkin, K.E., Camacho, E. and Osorio, J.E. 2016, Characterization of lethal Zika virus infection in AG129 mice, *PLoS Negl. Trop. Dis.*, **10**, e0004682.
51. Rossi, S.L., Tesh, R.B., Azar, S.R., . 2016, Characterization of a novel murine model to study Zika virus, *Am. J. Trop. Med. Hyg.*, **94**, 1362–9.
52. Sumathy, K., Kulkarni, B., Gondu, R.K., . 2017, Protective efficacy of Zika vaccine in AG129 mouse model, *Sci. Rep.*, **7**, 46375.
53. Brault, A.C., Domi, A., McDonald, E.M., . 2017, A Zika vaccine targeting NS1 protein protects immunocompetent adult mice in a lethal challenge model, *Sci. Rep.*, **7**, 14769.
54. Belshaw, R., Gardner, A., Rambaut, A. and Pybus, O.G. 2008, Pacing a small cage: mutation and RNA viruses, *Trends Ecol. Evol.*, **23**, 188–93.
55. Sabath, N., Wagner, A. and Karlin, D. 2012, Evolution of viral proteins originated de novo by overprinting, *Mol. Biol. Evol.*, **29**, 3767–80.
56. Goz, E., Zur, H. and Tuller, T. 2017, Hidden silent codes in viral genomes. In: Childs J.E., Mackenzie J.S., Richt J.A. eds., *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*, Springer: Berlin, Heidelberg, pp. 87–110.
57. Karlin, S., Burge, C. and Campbell, A.M. 1992, Statistical analyses of counts and distributions of restriction sites in DNA sequences, *Nucl. Acids Res.*, **20**, 1363–70.
58. Tulloch, F., Atkinson, N.J., Evans, D.J., Ryan, M.D. and Simmonds, P. 2014, RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies, *Elife*, **3**, e04531.
59. Mueller, S., Coleman, J.R., Papamichail, D., . 2010, Live attenuated influenza virus vaccines by computer-aided rational design, *Nat. Biotechnol.*, **28**, 723–6.
60. Martrus, G., Nevot, M., Andres, C., Clotet, B. and Martinez, M.A. 2013, Changes in codon-pair bias of human immunodeficiency virus type 1 have profound effects on virus replication in cell culture, *Retrovirology*, **10**, 78.
61. Haq, I.U., Chaudhry, W.N., Akhtar, M.N., Andleeb, S. and Qadri, I. 2012, Bacteriophages and their implications on future biotechnology: a review, *Virol. J.*, **9**.
62. Clark, J.R. and March, J.B. 2006, Bacteriophages and biotechnology: vaccines, gene therapy and antibacterials, *Trends Biotechnol.*, **24**, 212–8.
63. Hermoso, J.A., García, J.L. and García, P. 2007, Taking aim on bacterial pathogens: from phage therapy to enzybiotics, *Curr. Opin. Microbiol.*, **10**, 461–72.
64. Bull, J., Cunningham, C., Molineux, I., Badgett, M. and Hillis, D. 1993, Experimental molecular evolution of bacteriophage T7, *Evolution*, **47**, 993–1007.