Universal evolutionary selection for high dimensional silent patterns of information hidden in the redundancy of viral genetic code

Eli Goz\textsuperscript{1,2}, Zohar Zafrir\textsuperscript{1,2}, Tamir Tuller\textsuperscript{1,2,3,*}

\textsuperscript{1}Department of Biomedical Engineering, Tel-Aviv University, Ramat Aviv, Israel. \textsuperscript{2}SynVaccineLtd. Ramat Hachayal, Tel Aviv, Israel. \textsuperscript{3}Sagol School of Neuroscience, Tel-Aviv University, Ramat Aviv, Israel.

*Theses authors contributed equally to this work. *Corresponding author: tamirtul@post.tau.ac.il (TT)

Supplementary Information

1. Materials and methods

1.1 Virus-host database

The virus-host database table contains the raw data of the virus–host associations analyzed in this study (see Supplementary Table ST7): virushostdb.

The associations of viruses to their host organisms was retrieved from the GenomeNet Virus-Host Database (Mihara et al., 2016)(virus-host DB) that organizes this data in the form of pairs of NCBI taxonomy IDs. Virus-Host DB covers viruses with complete genomes stored in NCBI/RefSeq and GenBank, whose accession numbers are listed in EBI Genomes. Host information was collected from RefSeq, GenBank UniProt, and ViralZone, and manually curated with additional information obtained by literature surveys (about 38% of the total viral entries in the database are manually curated).

We first downloaded the virus-host database as tab separated file with tax ID, name, lineage and RefSeq IDs of a virus, and tax ID, name and lineage of its hosts ("Virus-Host Database,” n.d.). In case of segmented viruses, one entry may contain several different RefSeqIDs. In some cases one virus taxId may also contain several different RefSeqIDs corresponding to different version of a complete genome.

Coding Sequences and corresponding proteins of host organisms presenting in virus-host db were downloaded from Ensembl collections ("Ensembl Bacteria genomes collection,” n.d., “Ensembl Fungi genomes collection,” n.d., “Ensembl genomes collection,” n.d., “Ensembl Metazoan genomes collection,” n.d., “Ensembl Plants genomes collection,” n.d., “Ensembl Protists genomes collection,” n.d.)

For each downloaded host, "valid" coding sequences of the associated viruses were downloaded from ("Batch Entrez,” n.d.), according to their refSeqIds given in virus-
host database. The validity of coding sequences was asserted by comparing with corresponding translation products if given, if not sequences containing inframe stop codons in interior positions and/or a fractional number were omitted. In this way we assured that our results may be minimally affected by corrupted coding sequences.

In total we collected 2,625 unique viruses comprised of 147,286 coding sequences and mapped to 439 unique hosts (Figure 1B).

Figure S1: Viruses-hosts dataset summary. We analyzed 2,625 unique viruses belonging to different Baltimore classes: 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) viruses. These viruses were associated to 439 hosts from different domains of life: vertebrates, metazoa, plants, protists, fungi and bacteria (see methods). The top panel (grey bars) summarize the total number of viruses that infect at least one organism in each host domain (there may be viruses that infect organisms from different domains, e.g., arboviruses); the middle panel (color bars) specifies for each domain the portion of corresponding viruses belonging to each viral class; the bottom panel (grey bars) specifies the total number of different organisms in each domain.
1.2 Average host-repetitive and virus-repetitive substrings score (AHRS/AVRS).

The AHRS/AVRS scores are based on the tendency of substrings in a viral coding sequence $S$ to repeat in either a reference set $H$ comprised of coding sequences of the corresponding host (AHRS) or in a reference set $V$ comprised of the coding sequences of the same virus excluding the analyzed one (AVRS). They are defined as follows:

1) For each position $i$ in the coding sequence $S$ find the longest repetitive substring $S_i$ that starts in that position, and also appears at least once in $H$ (for AHRS) or in $V$ (for AVRS). In case of AVRS, common substrings found in the overlap regions of two coding sequences where excluded (this genomic overlap may be due to different mechanisms of coding capacity enhancement common in viruses, such as: alternative splicing, frameshifts, overlapping reading frames, etc.)

2) The AHRS/AVRS of sequence $S$ is the average length of all the substrings $S_i$

These scores are inspired by information theoretic approaches for universal compression of Markovian sequences, and estimating the number of bits required for describing one sequence ($S$) given a second reference sequence ($G$) (A.J.Wyner, 1993; Farach et al., 1994; Ziv and Lempel, 1977). More specifically, let $x^n$ denote a codon sequence of length $n$, and $M_s$ and $M_G$ stand for probability distributions of Markovian processes that generate codon distribution in $G$ and $S$ ($M_s(x^n)$ and $M_G(x^n)$ are the probabilities of emitting $x^n$ based on the corresponding Markovian models). Then, an average repetitive substring score of $S$ with respect to $G$ estimates the following measure:

$$\log(|G|) / - \sum_{x^n} M_S(x^n) \log \left( \frac{1}{M_G(x^n)} \right)$$

If the distribution of $S$ and $G$ are similar, $S$ can be better compressed by $G$. If $M_s = M_G$, the AHRS(AVRS) (the first equation) converges to

$$\log(|G|) / H(M_S)$$

where $H(M_S)$ is the entropy of $M_S$ and it is known that $H(MS)$ is smaller than $- \sum_{x^n} M_S(x^n) \log \left( \frac{1}{M_G(x^n)} \right)$ (second equation) for $M_s \neq M_G$. Thus theoretically longer genomes tend to have higher scores, while less "ordered" genomes (genomes with higher entropy) are characterized by lower scores.

The preprocessing step our approach is based on building a suffix array (Manber and Myers, 1993); this can be done in $O(|G|)$ where $|G|$ is the total length of the reference coding sequences (Gusfield, 1997). Then, the length of the longest substring starting...
at each position in a coding sequence that appears in the reference genome can be found in an efficient manner in $O(|S|)$. Thus, the total time complexity of the algorithm is $O(|G| + |S|)$.

The AHRS and AVRS scores were computed for each coding region individually and also once for each virus globally: a virus-specific AHRS score was computed by first excluding coding regions that are suspicious to be homologous to the host (see above) and then combining all the remaining sequences of the virus and their randomized variants. The virus-specific AVRS score was computed by averaging sequence-specific AVRS scores for all available coding regions.

If a virus is related to more than one host AHRS scores were computed for each host separately.

![Figure S2: Average host (virus) – repetitive scores](image)

1.3 Randomization models

We used the following two randomization models (see also Figure S3):

1. Dinucleotide Randomization (DNTR) - To preserve both the amino acids order and content, and the frequencies distribution of 16 possible pairs of adjacent nucleotides (dinucleotides) a model based on a multivariate Boltzmann sampling scheme was used (Zhang et al., 2013). This model was initially introduced in the context of enumerative combinatorics and was used by us before for studying synonymous information in specific viruses (Goz et al., 2017; Goz and Tuller, 2016, 2015). It produces random variants which feature both correct dinucleotide frequencies and coding capacity while being generated with provably uniform probability. We adapted the original source code which can be found in http://csb.cs.mcgill.ca/sparcs (“SPARCS webpagge,” n.d.).
(2) Synonymous Codon Randomization (SCDR) - To preserve both the *amino acids* order and content and the *codon usage* bias we used a Markov chain Monte Carlo method that generates a randomized sequence by iteratively swapping synonymous codons that encode the same amino acid.

![Dinucleotides randomization](image1.png)

![Synonymous codons randomization](image2.png)

**Figure S3: Randomization models**

### 1.4 Statistical analysis

The empirical p-values and z-scores, unless stated otherwise, were drawn from the empirical null distribution generated by the randomization models (see above); the p-value estimates the probability to get in random a value that is the same as, or more extreme than the observed result. The z-score estimates how far the observed result is from the mean value in standard deviation units derived from the null distribution:

\[
    p\text{-value} = \frac{|\text{rand}_{i} : \text{rand}_{i} \geq x|}{N+1} \\
    z\text{-score} = \frac{x - \text{mean}_{\text{rand}_{i}}}{\text{std}_{\text{rand}_{i}}}
\]

### 1.5 Analysis of low-dimensional features

The low-dimensional features were computed as follows:

**Effective Number of Codons (ENC)** is a measure that quantifies how far the codon usage of a coding sequence departs from equal usage of synonymous codons (Wright, 1990). For each amino acid (AA) let us define \(x_i\) to be the number of its synonymous codons of each type in the sequence, and \(n\) to be the number of times this AA appears in the sequence:

\[
    n = \sum_{i} x_i
\]

The frequency of each codon is therefore:

\[
    p_i = \frac{x_i}{n}
\]
The ENC for a specific AA is:

\[ N_e = \frac{1}{\hat{F}} \], where \( \hat{F} = \sum d p_i^2 \)

ENC for the group of AAs with degeneracy \( d \):

\[ N = \frac{1}{\hat{F}_d} \], where \( \hat{F}_d = \frac{1}{|A_d|} \sum_{i \in A_d} \hat{F}_i \)

(when an AA is missing, the corresponding effective number of codons is defined as an average over the given AAs of the same degeneracy).

Finally ENC for a virus is defined as an average of the group ENCs over all degeneracy AA groups weighted by the number of AAs in each group:

\[ N_e = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6} \]

computed over all viral coding sequence.

ENC can take values from 20, in the case of extreme bias where one codon is exclusively used for each amino acid (AA), to 61 when the use of alternative synonymous codons is equally likely. Therefore smaller ENC values correspond to a higher bias in synonymous codons usage; consequently, a negative correlation with ENC values means is equivalent to a positive correlation with synonymous codons usage.

**Codon Pairs Bias (CPB)**. To quantify codon pair bias, we follow (Coleman et al., 2008) and define a codon pair score (CPS) as the log ratio of the observed over the expected number of occurrences of this codon pair in the coding sequence. To achieve independence from amino acid and codon bias, the expected frequency is calculated based on the relative proportion of the number of times an amino acid is encoded by a specific codon:

\[
CPS = \log \left( \frac{F(AB)}{F(A) \times F(B) \times F(XY)} \right),
\]

where the codon pair AB encodes for amino acid pair XY and F denotes the number of occurrences. The codon pair bias (CPB) of a virus is then defined as an average codon pair scores over all codon pairs comprising all viral coding sequences:

\[
CPB = \frac{1}{k - 1} \sum_{i=1}^{k-1} CPS[i]
\]
Dinucleotide bias (DNTB) and CpG content. Following (Karlin, 1998) we compute a dinucleotide score (DNTS) for a pair of nucleotides XY as an odds ratio:

\[
DNTS = \frac{F(XY)}{F(X)F(Y)},
\]

where \(F\) denotes the frequency of occurrences.

Specifically, the CpG score is equal to the DNTS corresponding to the CG nucleotide

The dinucleotide pair bias (DNTB) of a virus is defined as an average of dinucleotide scores over all dinucleotides comprising all viral sequences:

\[
DNTB = \frac{1}{k-1} \sum_{i=1}^{k-1} DNTS[i]
\]

GC content is defined as:

\[
GC\% = \frac{F(G) + F(C)}{F(A) + F(G) + F(C) + F(T)}
\]

Where \(F()\) is a number of occurrences of each one of nucleotides A,G,C, and T.

1.6 Classification of viral genes into functional groups.

Viral coding regions were classified to 5 mutually exclusive functional groups: surface genes, structural genes, enzymes, others (accessory/regulatory) according to the properties encoded by them proteins. The group of each gene was determined by analyzing the annotations in related fasta headers according to a short – list of functional semantic keywords collected from a comprehensive literature survey; In additional, to improve the precision of our classification we used basic semantic relations between the keywords. For example: annotation containing an enzyme/surface keyword was classified as enzyme even if keywords from other structural groups appeared; annotations containing hypothetical keywords and keywords from other groups were assigned to the corresponding group (not to hypothetical group). Finally, the classification results were manually reviewed.

Examples of semantic keywords used for classification of coding regions into functional groups:

**Surface keywords:** recognition, receptor, surface, membrane, spike, glycoprotein, envelope, env, hn, hemagglutinin, fusion protein
**Structural keywords:** capsid, coat, core, matrix, structural protein, virion protein, attachment protein, capsomer, tegument, nucleoprotein, packaging protein, gag, pol, tail protein, head protein, neck protein, portal protein, binding protein, tape measure protein, head-tail joining protein

**Enzymes keywords:** enzyme names ending with the "ase" suffix

**Hypothetical proteins keywords:** hypothetical protein, putative protein, predicted protein

### 1.7 Finding enriched sequence motifs

Based on the significantly long substring sequences we identified, we looked for enriched *de-Novo* motifs using the HOMER (Hyper-geometric Optimization of Motif EnRichment) tool (Heinz et al, 2010).

Briefly, HOMER is designed for finding 8-20bp motifs in large scale genomics data, and is based on a differential motif discovery algorithm, *i.e.* it takes two sets of sequences and tries to identify the regulatory elements that are specifically enriched in one set relative to the other. It uses ZOOPS scoring (Zero or One Occurrence Per Sequence) coupled with the hyper-geometric enrichment calculations (or binomial) to determine motif enrichment.

The distance between two arbitrary motifs (*mot*$_1$ and *mot*$_2$), and between motif and TFBS/RBP was determined by comparison of the probability matrices using the following formula, which manages the expectations of the calculations by scrambling the nucleotide identities as a control (*freq$_1$* and *freq$_2$* are the matrices for *mot*$_1$ and *mot*$_2$, respectively):

\[
\text{Similarity Score} = \frac{1}{\text{Motif Length}} \sum_{i}^{\text{Motif Length}} \frac{(\text{Observed}_i - \text{Expect}_i)}{\text{Expect}_i}
\]

\[
\text{Observed}_i = \sum_{j}^{A,C,G,T} - (freq_1^{ij} - freq_2^{ij})^2
\]

\[
\text{Expect}_i = \sum_{j}^{A,C,G,T} \sum_{k}^{A,C,G,T} - \left(\frac{freq_1^{ij} - freq_2^{ik}}{4}\right)^2
\]

Neutral frequencies (0.25) are used in where the motif matrices do not overlap. The output score ranges from some lower bound (depending on the matrix frequencies) to 1, where 1 is complete similarity.
2. Results

2.1 Overview of the study

The general stages of our study appear in Figure S4: First, datasets of most of the known, as to the date of this study, virus-host associations were retrieved from (Mihara et al., 2016); available coding sequences of 2,625 unique viruses and 439 corresponding host organisms specified in this dataset were downloaded and preprocessed (Figure S1, Figure S4I-III). In order to demonstrate the evolutionary selection for long/complex patterns of silent functional information captured by AHRS/AVRS measures, we compared the wildtype viral sequences to 1,000 corresponding randomized variants (Figure S4IV). Two different randomization models that control for different mutational and selection biases were employed (Figure S2): the first, Synonymous Codon Randomization (SCDR), which preserves both the amino acid content and order, and the synonymous codon usage; the second, Dinucleotide Randomization (DNTR), which preserves both the amino acid content and order, and the frequencies of all possible dinucleotides (pairs of nucleotides). In addition, these randomization models preserve such basic sequence features as the encoded proteins and the frequencies of amino acids, codons, and mono and dinucleotides; however, they do not preserve more complex compositional patterns. If, indeed, there was a selection for a common high dimensional information that could not be explained merely by the amino acid content and order, nucleotide composition (e.g., GC content), preference for nucleotide pairs (e.g., CpG suppression) and codon usage bias (e.g., translation pressure on tRNA-codon affiliations), then we would expect longer or more abundant substrings of viral nucleotides to be repeated in the host or in the virus itself rather than in the corresponding randomized variants; respectively the AHRS and/or AVRS scores are expected to be higher in the wildtype than in random.

At the first step we analyzed the AHRS scores for each virus-host pair independently (one virus can have several hosts): in order to make sure that host-specific information reflected by AHRS can’t be attributed only to sequence similarity due to host-virus or virus-host horizontal gene transfer, viral sequences coding for proteins that are suspected to be homologous to at least one protein of the specific related host where excluded from the subsequent statistical analysis (Figure S4V). We then computed all host-repetitive substrings for all remaining real and randomized viral sequences with respect to the specific host (Figure S4VI). Consequently, sequence-specific AHRS scores and their empiric p-values with respect to both randomization models were computed for each viral coding region separately (Figure S4VII). In addition, a global virus-specific AHRS score and a corresponding p-value where computed globally for each virus by combining all its available sequences (that were not filtered out by host homology) (Figure S4VIII). Coding regions / viruses for which the sequence-specific / virus-specific AHRS scores were found to be significantly higher than in both randomizations models (p<0.05) were designated as AHRS - significant / selected for long host- repetitive substrings.
Significant coding regions were further analyzed in order to investigate whether the propensity to be selected for long host-repetitive substrings is related to the functional properties of the proteins encoded by them (Figure S4IX). Also in order to check whether certain sectors of a coding sequence tend to include longer repetitive sequences than others, local analysis of AHRS in 3 different parts of each coding sequence was performed (Figure S4X). In addition, relations between the global AHRS scores in AHRS-significant viruses and different low-dimensional features of their coding sequences, such as: Effective Number of Codons (ENC), Codon Pairs Bias (CPB), Dinucleotide Bias (DNTB), CpG and GC content, and the total length of coding sequences were examined (Figure S4XI).

At the second step, we analyzed the AVRS scores of a virus against itself (for viruses with at least two different coding sequences): for each viral coding sequence and its randomized variants, we filtered out its homologs appearing within the same viral genome (e.g., as a result of possible gene duplication events, gene transfer of similar sequences from the host, etc.; Figure S4XII), and computed all repetitive substrings with respect to the remaining coding sequences (excluding the analyzed sequence). To prevent the architecture of the viral genome from affecting the score, repetitive substrings found in overlapping parts of two coding sequences (e.g., due to alternative splicing, ribosomal frameshifts, overlapping reading frames, etc.) were omitted (Figure S4XIII). Then, sequence-specific and global virus-specific AVRS values, and their empiric p-values with respect to both randomization models were analyzed (Figure S4XIV-XV). As before, coding sequences / viruses for which the sequence-specific / virus-specific AVRS scores were found to be significantly higher than in both randomizations models (p<0.05) were designated as AVRS – significant / selected for long virus-repetitive substrings. Finally, we analyzed the tendency of viruses to be both AHR and AVRS significant (Figure S4XVI).
Figure S4: Flow diagram of the study.
2.2 AVRS analysis

We suggest that viral coding regions not only can contain patterns that are repeating in the coding regions of their hosts, but they also tend to include different local patterns that repeat in other coding regions of the same virus itself (Figure S5). Specifically, we found that such patterns are selected in the course of viral evolution in 47%, 46%, 27%, 50%, 33%, and 90% of viruses from different classes, that infect vertebrates, meatzoa, plants, protists, fungi, and bacteria correspondingly; they are on average significantly longer/more abundant (virus-specific AVRS p<0.05) than in random (i.e. we expect only 5% of viruses to be selected for by chance) and cannot be explained by the encoded peptides, compositional/mutational bias or by homologs and overlaps within the same viral genome. Distribution of corresponding significant virus-specific AVRS (and AHRS) scores is shown in Figure S6A-B.

It can be also seen, that the tendency of a virus to encode relatively long subsequences shared by its host (higher AHRS values than expected in random) and the selection for subsequences repeating in different coding sequences of the same virus (higher AVRS values than expected in random) are not mutually exclusive. In Figure S6C, we demonstrated that the portion of viruses that are both AHRS and AVRS significant is significantly (p<0.001) higher than expected in random (Figure S6D), for all host domains. On the other hand, we can see that host-repetitive and virus repetitive substring are not fully redundant, and one signal cannot be always explained by means of the other. Both of these evolutionary forces can often act both independently and together in the same virus, and both may have important roles in improving the viral fitness.
Figure S5: Selection for long virus–repetitive patterns of silent functional information in viral coding regions. Each vertical bar corresponds to viruses infecting a specific host organism (in bacteria – a specific genus) and is partitioned into class specific segments; every segment corresponds to percentage of viruses belonging to its corresponding class (y-axis) and is assigned a specific color. Further, each segment is composed of two stacked parts: the lower part with full color interior represents the portion (out of all host-specific viruses) of AVRS-significant viruses (p<0.05 w.r.t both randomization models); and the upper part with black interior (but with borders of the corresponding color) represents the rest of the viruses (p≥0.05 w.r.t at least one randomization model). Thus, for each class-specific segment, the sum of its two parts (significant and not significant) represent the total portion of viruses of this class within all viruses related to an organism described by the bar, and the sum of all segments is equal to 1. The horizontal bars visualizes the total percentage of AVRS-significant viruses in each host domain. We can see that coding regions in 47%, 36%, 39%, 27%, 25%, and 90% of viruses from different classes, that infect one or several vertebrates, metazoa, plants, fungi, protists, and bacteria organisms correspondingly, undergo an evolutionary pressure to maintain long genomic substrings that also tend to repeat in the other coding regions of the same virus.
Figure S6: A. Distribution of AHRS values for wildtype and randomized viral coding sequences. The average wildtype AHRS value (11.70) is higher than the average randomized (11.47) and this relation is statistically significant (Wilcoxon rank-sum right tail p<10^{-10}).

B. Distribution of AVRS values for wildtype and randomized viral coding sequences. The average wildtype AVRS value (8.45) is significantly higher than the average randomized (6.43) and this relation is statistically significant (Wilcoxon rank-sum right tail p<10^{-100}).

C. Percentage of viruses that are both AHRS and AVRS significant is higher than expected in random. The overlap between these two sets of viruses was measured (for each host domain separately) by their intersection over union (IoU) and

\[
\text{Intersection over Union (IoU)} = \frac{\text{%AHRS signif} \cap \text{%AVRS signif}}{\text{%AHRS signif} \cup \text{%AVRS signif}}
\]
compared with the values expected in random (see also Methods). **D. One-versus-rest randomization model for testing the statistical significance of the overlap between the sets AHRS and AVRS significant viruses.** The randomized intersection over union values were modeled as follows: for each randomized variant of each wild-type virus, we compared its AHRS (AVRS) values to the corresponding values of the remaining randomized variants of the same virus (each row in a column is compared with the rest of the rows in the same column). As a result, we obtained 1,000 sets of randomized AHRS (AVRS) p-values; each set (row) containing one randomized p-value for each one of the viruses. We then identified those variants with p-value < 0.05 (marked in yellow) and computed the intersection over union of these variants for AHRS and AVRS for each row, yielding 1000 randomized intersection over union values. This algorithm was performed separately for DNTR and SCDR randomization models; the results were unified and plotted in the histogram (blue). The wildtype IoU values for each host kingdom are plotted by color lines. We can see that that the portion of viruses that are both AHRS and AVRS significant is significantly (p<0.05) higher than expected in random in all host domains.

2.3 The long host-repetitive silent patterns cannot be explained only by low dimensional genomic features

As was previously mentioned, various basic characteristics of viral genomes may be related to the viral fitness and life cycle. In order to analyze the relations of such characteristics to the selection for more complex / long host – repetitive silent patterns reported here, we computed, for each virus, the following 'low-dimensional' genomic features (LDF): Effective Number of Codons (ENC), Codon Pairs Bias (CPB), Dinucleotides Bias (DNTB), GC and CpG content and the total length of (non-host homologous) coding sequences (details in the Methods section). The results of comparison of these features to the corresponding combined z-scores of AHRS values (AHRS-z) averaged across the different random models are shown in Figure S7. We can see that for AHRS-significant viruses, the Spearman rank correlation between the LDF and AHRS-z residuals (the variation in LDF and AHRS-z variables after controlling for the length of the sequences) cannot explain the selection for long-host repetitive patterns (Figure S7A) merely by their relation to more basic genomic features. The negative correlation with ENS (in all domains, but fungi) can be explained by a selection pressure on adaptation to the host which can be manifested both by adaptation of codons and by longer/more complex patterns. The negative correlation with CpG in vertebrates and the positive correlation in bacteria may be a consequence of the fact that CpG pairs are suppressed in the former and prevalent in the latter genomes (Cooper and Krawczak, 1989; Krieg, 2002). Also, a positive correlation with GC content in all host kingdoms with the strongest one in bacteria was observed; this may be related to the fact that in some cases highly expressed genes tend to have stronger mRNA folding and thus GC content (due to strong relation between GC content and folding strength (Zur and Tuller, 2012)). The correlation with Codon Pairs Bias is very small (0.01-0.16) and is negative or close to zero in all hosts but fungi; this suggests that in general the results reported here do not strongly overlap with Codon Pairs Bias.
In addition we found no significant differences between the LDF and AHRS-z residuals corresponding to two groups of viruses: (i) AHRS-significant and (ii) – AHRS not significant (Figure S7B).

### Figure S7: Relation between low dimensional features and AHRS z-scores.

LDF values and corresponding AHRS-z scores (AHRS-z) may be significantly correlated (either positively or negatively) due to a positive correlation of both of them with the genome length. Therefore, to control for the viral genome length, we computed a partial correlation between the LDF and combined AHRS Z values (computed as an average of the AHRS-z scores with respect to each random model). A partial correlation between X and Y, given a control variable Z, is the correlation between the residuals R_X and R_Y resulting from the linear regression of X with Z and of Y with Z respectively. These residuals are actually the variation in X and Y variables that are not explained by the control variable Z. In the figure, the length-controlled relations between LDF values of AHRS- significant and not-significant viruses, and between the AHRS-z and different LDFs for significant viruses are demonstrated for each

| Feature       | LDF Values | AHRS-z Values |
|---------------|------------|---------------|
| ENC resid.    |            |               |
| CPB resid.    |            |               |
| DNTB resid.   |            |               |
| CpG resid.    |            |               |
| GC resid.     |            |               |
| log(length)   |            |               |

- **Figure S7A**: Scatter plots showing the correlation between LDF values and AHRS-z scores for different viral groups.
- **Figure S7B**: Summary of p-values indicating the significance of correlations.
LDF and host domain (due to a small number of protest viruses, this domain was excluded). The LDF and AHRS-z variables are represented by their Z-standardized (mean 0, variance 1) residuals (in the analysis of AHRS-z with genome length itself, a regular spearman correlation was performed, since in this case the correlated variable is the same as control). A. A scatter plot LDF residues and AHRS-z residues with corresponding partial Spearman correlation values and least square lines. B. A comparison of LDF residues for AHRS-significant and not-significant viruses.

Finally, in order to further demonstrate that the observed patterns of significantly long substrings cannot be entirely explained by simple characteristics (i.e. LDFs) of the genomic sequences, we performed a regression analysis on all of the low dimensional features. To this end, we build a linear regression model that uses all these LDFs and is aimed to maximize the correlation with the AVRS/AHRS scores. The regression model was separately performed on each host group with more than 100 significant viruses (i.e. separately for Bacteria, Plants, and Vertebrates). For each host group, we randomly separated all the 6 features into two sets using 50% of the viruses as a train set and the rest (50%) as a test set (i.e. each contains 50% of the viruses having a significant AVRS score; see more details in the supplementary data). Hence, the regression model was build based on the first group and was tested on the second group, by performing a correlation between the predicted and the actual AVRS scores. Results gave regression correlation of 0.39<r<0.71, when using all the 6 features (p<4.58·10^{-7}); see details in Supplementary Table ST2). This demonstrates that only up to 15%-50% of the variance can be explained by these ‘low dimensional’ features. Furthermore, the results of comparison of these features to the AHRS statistics of the corresponding genomes, demonstrated explicitly that selection for long host-repetitive patterns cannot be explained merely by their relation to more basic genomic features. Thus, we conclude that the low dimensional features typically explain relatively low percentage of the AVRS score variability.

2.4 Long substrings analysis

The coding regions of many viruses from all classes that infect different organisms from all domains of life tend to undergo evolutionary selection for long patterns of silent functional information that may be important to their fitness. These patterns are encoded in viral genomic substring repeats in the coding regions of viruses and in the coding regions of their hosts. In order to further understand how these patterns properties, we generated distribution histograms of the substring length. Specifically, we considered only substrings that were significantly longer than expected (i.e. with p<0.05 compared to our randomized models). The median substring length was found to be 39 (Figure S8A). Additionally, we separated the substrings into various subgroups according to their host type (Fungi, Bacteria, Plants, Vertebrates, Protists, and Metazoa) and virus type, based on the Baltimore classification (Figure S8B-C); see details in Supplementary Table ST1.
Figure S8: Analysis of length distribution of significantly enriched substrings. A. All Viruses. B. Division according to host type. C. Division according to virus type.
2.5 Sequence enrichment analysis: de-novo sequence motifs, transcription factors, and RNA binding proteins in human viruses

In order to further understand how the repetitive patterns found promote viral fitness and affect gene expression, we performed comprehensive inspection and looked for de-novo sequence motifs that appear in the repetitive substrings of human viruses. Specifically, we analyzed these significantly long substrings using an algorithm for finding de-novo sequence motifs (Heinz et al., Mol Cell, 2010) that appear in human viruses more than in comparison to our randomized models; see also previous sections. The analysis found 1125 significant motifs (p<0.05); the motifs were sorted by their p-values, and after controlling for false discovery rate (FDR; q=0.01), we end up with 1089 motifs; see Supplementary Table ST3. Similar motifs, with similarity score higher than 0.6 can be found in Supplementary Table ST4.

Next, we compared these motifs against known information of transcription factor biding sites (TFBS) and RNA biding proteins (RBPs), which were taken from the JASPAR (Khan et al., NAR, 2018) and RBPmap (Paz et al., NAR, 2014) databases. Specifically, for each host-virus pair we used these substrings as a target set, and compared them to a similar background set of substrings, taken from our randomized models (i.e. sub-sequences with the same length, GC content, CUB, etc.). The results show enrichments of TFs related to the following classes: Basic helix-loop-helix factors (bHLH), C2H2 zinc finger factors, and Tryptophan cluster factors. We also found enrichments of RBPs for the HNRNPxx, PABPxx, and SRFSx proteins; see more details in Supplementary Tables ST5-ST6.

Finally, we performed this type of analysis on target sets of substrings and on randomized viral genomes, which maintain the encoded protein sequences, the codon frequencies and GC content (but not the codons order); this demonstrated that we get significantly lower number of TFs/RBPs in this cases in comparison to the analysis done on the actual data (p<0.03 and p<0.04, respectively). This result supports our hypothesis that indeed evolution shape viral coding region to include "meaningful" sub-sequences, longer than single codons, important to the viral fitness, and also provides an interesting explanation regarding the function of some of the detected sub-sequences.
2.6 Additional analysis of the dependence of enrichment for AHRS significant coding regions in different functional proteins groups and their length

The purpose of this section is to demonstrate that the reported results related to the enrichment of structural proteins (and relative to other protein groups), are not due to different typical length of structural proteins, in comparison to other proteins.

To this end, we separated the genes into 4 groups according to their ORF’s length (see Table S1 bellow), and shows that we still get higher and significant enrichment in the structural proteins.

| Gene groups | Lengths of analyzed genes | ≤ 500 | 501-1000 | 1001-1500 | >1500 |
|-------------|----------------------------|-------|----------|-----------|-------|
| surface     | # genes                   | 666   | 570      | 264       | 397   |
|             | %signif                   | 0.08  | 0.11     | 0.11      | 0.24  |
|             | mean ORF length           | 339   | 790      | 1166      | 2494  |
| structural  | # genes                   | 3589  | 3735     | 2211      | 3525  |
|             | %signif                   | 0.1   | 0.25     | 0.34      | 0.38  |
|             | mean ORF length           | 378   | 773      | 1233      | 2674  |
| enzymes     | # genes                   | 5418  | 6006     | 3355      | 4066  |
|             | %signif                   | 0.05  | 0.15     | 0.20      | 0.27  |
|             | mean ORF length           | 394   | 742      | 1242      | 2327  |
| hypothetical| # genes                   | 68584 | 14937    | 3205      | 2461  |
|             | %signif                   | 0.09  | 0.22     | 0.31      | 0.38  |
|             | mean ORF length           | 315   | 699      | 1199      | 2538  |
| unclassified| # genes                   | 27933 | 10347    | 3803      | 4133  |
|             | %signif                   | 0.1   | 0.25     | 0.30      | 0.35  |
|             | mean ORF length           | 322   | 730      | 1216      | 2771  |

Table S1: The analyzed viral coding regions where divided into 4 bins according to their lengths. The enrichment of AHRS significant genes in different functional proteins groups was analyzed for each gene independently. We can see that within each bin the structural group is the most enriched one and the surface group and enzymes are less enriched. Therefore our conclusions cannot be attributed to the lengths of coding regions.
3. Discussion

Figure S9: Two suggested non-mutually exclusive hypotheses related to the observed tendency of viruses to include long sub-sequences/codes in their coding regions that appear also in the host. A. These codes enable efficient immune system avoidance; specifically they may enable escaping the CRISPR system. B. The codes enable a better adaptation to the host gene expression machinery.
4. References

A.J. Wyner, 1993. String Matching Theorems and Applications to Data Compression and Statistics. Ph.D Thesis, Stanford Univ.
Batch Entrez [WWW Document], n.d. URL http://www.ncbi.nlm.nih.gov/sites/batchentrez
Coleman, J.R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., Mueller, S., 2008. Virus attenuation by genome-scale changes in codon pair bias. Science 320, 1784–7.
Cooper, D.N., Krawczak, M., 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. 83, 181–8.
Ensembl Bacteria genomes collection [WWW Document], n.d. URL http://bacteria.ensembl.org/index.html
Ensembl Fungi genomes collection [WWW Document], n.d.
Ensembl genome collection [WWW Document], n.d. URL http://www.ensembl.org/info/about/species.html#eg
Ensembl Metazoa genomes collection [WWW Document], n.d. URL http://metazoa.ensembl.org/index.html
Ensembl Plants genomes collection [WWW Document], n.d. URL http://plants.ensembl.org/species.html
Ensembl Protists genomes collection [WWW Document], n.d. URL http://protists.ensembl.org/index.html
Farach, M., Farach, M., Noordewier, M., Savari, S., Shepp, L., Wyner, A., Ziv, J., 1994. On the Entropy of DNA: Algorithms and Measurements based on Memory and Rapid Convergence. Proc. SIXTH Annu. ACM-SIAM Symp. Discret. ALGORITHMS 48–57.
Goz, E., Mioduser, O., Diament, A., Tuller, T., 2017. Evidence of translation efficiency adaptation of the coding regions of the bacteriophage lambda. DNA Res.
Goz, E., Tuller, T., 2015. Widespread signatures of local mRNA folding structure selection in four Dengue virus serotypes. BMC Genomics 16 Suppl 1, S4.
Goz, E., Tuller, T., 2016. Evidence of a Direct Evolutionary Selection for Strong Folding and Mutational Robustness Within HIV Coding Regions. J. Comput. Biol. 23, 641–650.
Gusfield, D., 1997. Algorithms on strings, trees, and sequences: computer science and computational biology. Cambridge University Press.
Heinz, S., et al., 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities, Molecular Cell, 38, 576-589.
Karlin, S., 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. Curr. Opin. Microbiol. 1, 598–610.
Khan, A., et al., 2018. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 46, D260-D266.
Krieg, A.M., 2002. CpG motifs in bacterial DNA and their immune effects. Annu. Rev. Immunol. 20, 709–760.
Manber, U., Myers, G., 1993. Suffix Arrays: A New Method for On-Line String Searches. SIAM J. Comput. 22, 935–948.
Mihara, T., Nishimura, Y., Shinizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., Ogata, H., 2016. Linking Virus Genomes with Host Taxonomy. Viruses 8, 66.
Paz, I., Kosti, I., Ares, M., Cline, M., Mandel-Gutfreund, Y., 2014. RBMap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res. 42, W361-W367.
SPARCS webpage [WWW Document], n.d. URL http://csb.cs.mcgill.ca/sparcs/
Virus-Host Database [WWW Document], n.d. URL http://www.genome.jp/virushostdb/
Wright, F., 1990. The “effective number of codons” used in a gene. Gene 87, 23–9.
Zhang, Y., Ponty, Y., Blanchette, M., Lécuyer, E., Waldispühl, J., 2013. SPARCS: a web server to analyze (un)structured regions in coding RNA sequences. Nucleic Acids Res. 41, W480-5.
Ziv, J., Lempel, A., 1977. A universal algorithm for sequential data compression. IEEE Trans. Inf. Theory 23, 337–343.
Zur, H., Tuller, T., 2012. Strong association between mRNA folding strength and protein abundance in S. cerevisiae. EMBO Rep. 13, 272–7.