Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

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

Eli Goz[1,2,&], Zohar Zafrir[1,2,&], Tamir Tuller [1,2,3*]

[1]Department of Biomedical Engineering, Tel-Aviv University, Ramat Aviv, Israel. [2]SynVaccineLtd. Ramat Hachayal, Tel Aviv, Israel. [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)

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Understanding how viruses co-evolve with their hosts and adapt various genomic level strategies in order to ensure their fitness may have essential implications in unveiling the secrets of viral evolution, and in developing new vaccines and therapeutic approaches. Here, based on a novel genomic analysis of 2,625 different viruses and 439 corresponding host organisms, we provide evidence of universal evolutionary selection for high dimensional 'silent' patterns of information hidden in the redundancy of viral genetic code. **Results:** Our model suggests that long substrings of nucleotides in the coding regions of viruses from all classes, often also repeat in the corresponding viral hosts from all domains of life. Selection for these substrings cannot be explained only by such phenomena as codon usage bias, horizontal gene transfer, and the encoded proteins. Genes encoding structural proteins responsible for building the core of the viral particles were found to include more host-repeating substrings, and these substrings tend to appear in the middle parts of the viral coding regions. In addition, in human viruses these substrings tend to be enriched with motives related to transcription factors and RNA binding proteins. The host-repeating substrings are possibly related to the evolutionary pressure on the viruses to effectively interact with host's intracellular factors and to efficiently escape from the host's immune system.

Contact: : tamirtul@post.tau.ac.il (TT)

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Viruses are subcellular particles, consisting of encapsulated genomic material, that replicate only inside the living cells of other organisms. Being under permanent pressure to escape from the defense mechanisms of the cell and at the same time driven by an essential requirement to ensure optimal conditions for efficient and selective replication, viruses are forced to continuously co-evolve with the host by adapting various properties and mechanisms, often uncommon to cellular organisms (Domingo, n.d.; Firth and Brierley, 2012; Gale et al., 2000; Gibbs et al., 2005; Holmes and Drummond, 2007; López-Lastra et al., 2010). These mechanisms can involve the recruitment and/or modification of cellular factors, but are also inherent in the nucleotide composition of the viral genomic sequences themselves. In particular, viral genomes, and specifically coding regions, not only determine protein products, but also include additional, overlapping, information encrypted in the combination of synonymous codons. This information doesn't affect the protein encoding (*i.e.* phenotypically 'silent'), and is associated with different biophysical and evolutionary aspects related, among others, to the amplification of the genomic coding potential, to the regulation of viral gene expression, and to the mediation of intercellular interactions (Brierley, 1995; Cuevas et al., 2012; Firth and Brierley, 2012; Gale et al., 2000).

Accordingly, it is reasonable to anticipate that the genomic footprints of virus host co-evolution could be seen in the form of common compositional signatures shared both by viral and host genomes. Indeed, examination of such signatures has revealed correspondences between the genomes of viruses from different specific groups and their hosts (Barrai et al., 1990; Cardinale and Duffy, 2011; Greenbaum et al., 2008; Jenkins et al., 2001; Kerr and Boschetti, 2006; Lobo et al., 2009; Mihara et al., 2016; Pride et al., 2006; van Hemert et al., 2007). For example in (Bahir et al., 2009; Mihara et al., 2016) a significant correlation between GC content of bacteriophages with their prokaryotic hosts was demonstrated (although no significant associations were found for other taxonomic groups). In (Greenbaum et al., 2008; Lobo et al., 2009; Shackelton et al., 2006) it was shown that CpG pairs are under-represented in many RNA and most small human DNA viruses, in correspondence to dinucleotide frequencies of their hosts. Further motivated by the possibility that a complete dependence on the translational machinery of a cell might subject the codon

usage of viral genes to host selection pressures, various studies have focused on exploring the similarity between the codon usage preferences in viruses and their hosts. These studies revealed numerous examples of viral codon usage either matching or significantly deviating from the codons usage for corresponding organisms from different taxonomic domains (Bahir et al., 2009; Barrai et al., 2008; Carbone, 2008; Cheng et al., 2012; Coleman et al., 2008; Gu et al., 2004; Kunec and Osterrieder, 2016; Lobo et al., 2009; Lucks et al., 2008; Mueller et al., 2006; Sau et al., 2007, 2005; Su et al., 2009; Zhao et al., 2008). Nevertheless, almost all of the previous works examined a limited number of specific viral families and were mostly based on comparisons of very basic (low-dimensional) compositional characteristics of genomic sequences such as: GC content, dinucleotides, codons, and more generally short oligomers. Although these features may present, to some extent, evidence for possible virus-host co-adaptation, they cannot fully capture longer patterns of information. For example, transcription factors binding sites (TFBS), the binding site of micro-RNAs, RNA binding proteins (RBPs), spliceosome, sequences related to immune system (*e.g.*, CRISPR), etc., can typically be longer than 10nt and can vary among different viruses, cells and host organisms; therefore they cannot be fully described by simple features spanned by short nucleotide k-mers. Since viral genomes co-evolve with their hosts and adapt the function and expression of their genes to interact with intracellular environments, we expect such longer patterns to appear both in the viral and in the cellular coding regions and play role in controlling the viral fitness.

In this study we performed for the first time a large scale computational analysis of long patterns of silent functional information that repeat in coding regions of viruses and their associated hosts. Our analysis was based on the largest viral-host dataset analyzed so far that contains most of the available virus-host associations and covers 2,625 unique viruses of all classes and 435 different hosts from all kingdoms of life. We have shown that the coding regions of many viruses tend to undergo evolutionary selection for inclusion of repeating substrings that are on average longer or more abundant than expected in random, and cannot be explained by the encoded viral/host proteins or by basic genomic features such as the preferences for synonymous codons/codon pairs or distribution of nucleotide pairs. Nor can they be explained by gene transfer mechanisms or by canonical mechanisms of protein recognition by the immune system alone. Our approach was inspired by universal methods for data compression without any prior knowledge of its statistical characteristics (Ulitsky et al., 2006; Ziv and Lempel, 1977) and is based on the idea that various aspects of viral fitness are encoded in the composition of synonymous codons by possibly long patterns of nucleotides that tend to appear in the coding regions of both viral and host genomes. Our results provide evidence of a complex genomic level evolutionary adaptation of viruses to their hosts and may have important implications in understanding the viral evolution and developing novel antiviral vaccines and therapeutic approaches.

2 Methods

In this section we briefly summarize the most important rationale of our methodology. The details appear in **Supplementary sections 1.1-1.7**.

2.1 Data preparation

The associations of viruses to their host organisms was retrieved from the GenomeNet Virus-Host Database (virus-host DB) (Mihara et al., 2016). In total we collected 2,625 unique viruses comprised of 147,286 coding sequences and mapped to 439 unique hosts. To date, this is the largest virus-host analysis, based on most of the known virus-host associations reported (see also **Supplementary section 1.1**).

2.2 Average repetitive substring scores

We defined two types of scores called: average virus-repetitive substrings (AVRS), and average host-repetitive substrings (AHRS); as their names suggest, these scores quantify the average length of all possible substrings that repeatedly appear in the coding sequences of a virus itself, and/or in the coding sequences of its host (*i.e.* AVRS and/or AHRS, respectively). They are motivated by the assumption that evolution shapes the viral coding sequences to improve their interaction with the intra-cellular environment. Thus, if longer (than expected from compositional biases driven by neutral evolutionary forces) substrings of a coding region tend to appear also in host and/or other viral coding sequences, it may suggest that these substrings are associated with functional synonymous motives related to various aspects of viral replication and have been selected by evolutions to improve viral fitness, *e.g.*, via adaptation to the cellular gene expression machinery or to the innate immune system. These scores can potentially capture known and unknown (or hidden) high dimensional (longer than codon or short k-mers) information encoded in the genomic substrings of nucleotides of an arbitrary length. They can be efficiently and systematically applied to a large scale set of viruses and their related hosts in an unsupervised manner, *i.e.* without a prior knowledge on the intrinsic genomic structure shaped by these associations, and with no prior knowledge on the substring length. In addition, as was previously demonstrated in (Zafrir and Tuller, 2017; Zur and Tuller, 2015), such scores are able to capture complex information that does not appear in single codon/codon-pairs distributions and in particular to be used for predicting the expression levels/protein levels of a gene from its sequence.

The AVRS/AHRS scores are computed as follows (see more details in **Supplementary section 1.2**): (1) Build a suffix array (Manber and Myers, 1993) – this can be done in O(|H(V)|) (Gusfield, 1997); (2) For each position *i* in a viral coding sequence S, use the suffix array from (1) to find the longest repetitive substring S_i that starts at that position, and also appears at least once in H(V) (for AVRS) – this can be done in in O (|S|). 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 the coding capacity enhancement common in viruses, such as: alternative splicing, frameshifts, overlapping reading frames, etc.); (3) The AVRS/AHRS of a sequence S is the average length of all the substrings S_i. The total time complexity of the algorithm is O(|H(V)| + |S|). The scores are computed for each viral coding sequence individually.

2.3 Sequence homology

In order to make sure that host-specific information reflected by AVRS/AHRS can't be attributed only to sequence similarity due to host-virus or virus-host horizontal gene transfer (HGT), as well as to repeats in viral genomes due to gene duplications or transfer of similar sequences from the host, viral sequences coding for proteins that are suspected to be homologous to at least one protein of the related host (virus-host homology), and/or to at least one other protein of the same virus (virus-virus homology), where excluded from the subsequent statistical analysis. To this end, we constructed a local BLAST (Altschul et al., 1990) database comprising all downloaded host/virus proteins. Each viral coding sequence was translated, and the resulting protein sequence was queried against the database of host/virus proteins. Any match within the proteome of the corresponding virus/host with e-value < 0.0001 was defined as homologous and the corresponding viral sequence was excluded from further analysis. We used BLAST version 2.4.0 (http://blast.ncbi.nlm.nih.gov).

2.4 Randomization models and statistical analysis

To test our hypothesis regarding the selection for longer repetitive substrings, we used the following two randomization models: (1) Dinucleotide Randomization that preserves both the amino acids order and content, and the frequencies distribution of 16 possible pairs of adjacent nucleotides (dinucleotides); (2) Synonymous Codon Randomization that preserves the amino acids order and content, mono-nucleotide composition, and the codon usage bias (see also **Supplementary section 1.3**).

If, indeed, there was a selection for high dimensional information patterns that could not be explained by the basic genomic features preserved in these models, then we would expect longer substrings of viral nucleotides to be repeated in the host or in the virus itself to a greater extent than in the corresponding randomized variants; respectively the AVRS/AHRS scores are expected to be higher in the wildtype than in comparison to randomized genomes.

Empirical p-values and Z-scores, unless stated otherwise, were drawn from the empirical null distribution generated by the above randomization models. 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 empirical z-score estimates how far the observed result is from the mean value in standard deviation units derived from the null distribution (see **Supplementary section 1.4**).

3 Results

3.1 Overview of the analysis

The general stages of our study are as follows (see more details in **Supplementary section 2.1** and **Supplementary Figure S4**): Virus-host data was downloaded and preprocessed. In order to demonstrate the evolutionary selection for long patterns of silent functional information captured by AVRS/AHRS measures, we compared the wildtype viral sequences to 1,000 corresponding randomized variants generated by each of the described above randomization models. We use the term "silent" patterns in this paper since the null model maintains the amino acid composition of the original encoded proteins in the virus. Thus, the AHRS/AVRS can be explained only by aspects of the coding sequence that are not related to the amino acid composition (*i.e.* 'silent').

First we analyzed the AHRS scores for each virus-host pair independently (one virus can have several hosts and vice versa): Consequently, sequence-specific AHRS scores and their empiric p-values and Z-scores with respect to both randomization models were computed for each viral coding region separately. In addition, a virus-specific AHRS scores and the corresponding p-values and Z-scores were computed globally for each virus by combining all its available coding sequences. 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, *i.e.* selected for long host-repetitive substrings. AHRS-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 corresponding proteins. Also in order to check whether certain sectors of a coding sequence tend to be enriched with longer host-repetitive sequences more than others, local analysis of AHRS in 3 different equal parts of each coding sequence was performed. In addition, explicit relations between the global AHRS scores in AHRS-significant viruses and different low-dimensional genomic features (LDF) 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. Finally, a similar analysis was performed to study the AVRS scores of a virus against itself (for viruses with at least two different coding sequences).

3.2 Evidence of universal selection for long patterns of silent functional information inside viral coding regions

Our analysis suggests that the coding regions of many viruses from all classes, which 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; these substrings are generally longer than a single codon, codon pairs, or short k-mers of nucleotides (median=39, for positions with $p<0.05$); see details in **Supplementary section 2.4** and **Supplementary Figure S8**. Furthermore, they cannot be entirely explained by simple characteristics (*i.e.* LDFs) of the genomic sequences (such as amino acids order and content, compositions of mono and di-nucleotides, codon bias, etc.). Specifically, a regression model taking into account a combination of these features demonstrates that only up to 15%-50% of the variance can be explained by them ($p<4.58\cdot10^{-7}$). 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 (see **Supplementary sections 1.5** and **2.3** for more details).

Specifically, we have found that many of the analyzed viruses and their hosts undergo significant enrichment for mutually long substring. Thus, more than 56% of the analyzed human viruses and 90% of the analyzed bacteriophages, undergo an evolutionary pressure to maintain genomic substrings that also tend to repeat in the coding regions of at least one related host (**Figure 1**). These substrings are apt to be on average significantly longer (virus-specific AHRS $p<0.05$) than expected if only lower-dimensional silent functional information was selected for (*i.e.*, we expect only 5% of viruses to be selected for by chance). The distribution of their corresponding virus-specific AHRS values is shown in **Supplementary Figure S6A**.

In a similar manner we demonstrated that viral coding regions not only contain patterns that are repeated in the coding regions of their hosts, but also tend to include silent local patterns that repeat in other coding regions of the *virus itself*. 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), are on average significantly longer (virus-specific AVRS $p<0.05$) than in random and cannot be explained by the encoded proteins, compositional / mutational bias or by homologs and overlaps within the same viral genome; see more details in **Supplementary section 2.2** and **Supplementary Figure S5**. Distribution of the corresponding virus-specific AVRS scores as well as additional analysis can be found in **Supplementary Figure S6B-D.**

3.3 Enrichment of *de-novo* sequence motifs, transcription factors, and RNA binding proteins found in human viruses

Following, and in order to further understand how the patterns found promote viral fitness, we performed comprehensive analysis of the significantly long substrings using an algorithm for finding *de-novo* sequence motifs (Heinz et al., 2014) that appear in human viruses more than expected by the our null model (see **Supplementary section 1.7**). Next, we compared these motifs against known information of TFBS and RBPs, taken from the JASPAR (Khan et al., 2014) and RBPmap (Paz et al., 2014) databases.

We found enrichment of transcription factors (TFs) related to the following classes: Basic helix-loop-helix factors (bHLH), C2H2 zinc finger factors, and Tryptophan cluster factors, and enrichment of RBPs for the HNRNPxx, PABPxx, and SRFSx proteins. We also find that generally these viral genomes tend to include more TF and RBS binding sites than expected from a Null model ($p<0.04$); see more details in **Supplementary section 1.8** and **Supplementary tables ST3-ST6**. This provides one interesting explanation regarding the function of some of the detected sub-sequences.

3.4 Selection for long host-repetitive silent patterns depends on the protein's function

The genomes of all known viruses encode structural proteins, which serve as building units of viral particles or are responsible for the interaction with the host receptors and invasion to the cell. In addition, most of the viruses express some replication enzymes, such as reverse transcriptase or RNA/DNA polymerase, according to their mode of replication, transcription, and regulation. The rest of the viral proteome is responsible for diverse regulatory/accessory functions, which are mostly uncharacterized and often specialized to the life cycle of the particular virus.

Here we aimed at refining the resolution of the genome level analysis previously presented, and to find out whether specific group of proteins is more favored by selection for long synonymous patterns than others. To this end, we classified the analyzed viral genes to 5 mutually exclusive functional groups (see also **Supplementary section 1.6**): surface genes, structural genes, enzymes, hypothetical (putative proteins), and unclassified (accessory or regulatory proteins). In **Figure 2** we show that 13%, 28%, 18%, 15%, 21% of the

coding sequences belong to surface genes, structural genes, enzymes, and genes corresponding to putative and unclassified proteins have significantly high sequence specific AHRS scores ($p < 0.05$ with respect to both randomization models). We can see that structural proteins that do not function as host recognition elements are characterized by the highest portion of AHRS significant genes (28%, Fisher exact test $p<1\cdot10^{-16}$). On the other hand, among proteins expressed on the viral surface, which participate in recognition of the host receptors and often susceptible to higher mutability, the number of AHRS significant genes is the smallest (13%). The enzymes (18%) and other unclassified proteins show an intermediate level of selection for long host-repetitive patterns.

In order to reinforce the claim that our conclusions can't be attributed only to sequence lengths, the analyzed viral coding regions were divided into 4 bins according to their length. The percentage of AHRS significant genes in different functional groups was analyzed for each bin independently. Again, we observed that within most of the bins the structural group contains the highest number of AHRS - significant genes, and the surface group and enzymes – the lowest number. Therefore, our conclusions cannot be attributed only to the lengths of the coding regions (see details in **Supplementary section 2.6**). This finding is in agreement with stronger codon usage resemblance of viral structural genes to their host sequences demonstrated in (Bahir et al., 2009), and may be attributed to higher expression levels required from this functional group. Thus, this group should be under stronger selection for optimal gene expression codes; the higher expression levels may also have stronger effect on the host immune system, triggering stronger selection to include longer pattern similar to the host.

Finally, we were interested in checking whether there is a preference for longer host-repetitive subsequences in specific parts of coding sequences. To this end, we divided each coding sequence into 3 equal parts, corresponding to the beginning, the middle and the end of the sequence, and calculated local AHRS scores inside each one of them. We found that for each gene group, most of the sequences used to have the highest local AHRS in the middle part; the percentage of genes with the highest local AHRS in the 3' part was found to be the smallest. This pattern may be related to the fact that initiation and termination (of translation and transcription), encoded in the coding region ends, tend to be non-canonical in viruses (*e.g.*, initiation via IRES), while the regulation at the middle of the coding region is more conserved relatively to the host (Clyde and Harris, 2006; Gale et al., 2000; Groat-Carmona et al., 2012; Jackson, 2005; Kieft, 2008; López-Lastra et al., 2010; Thurner et al., 2004). In addition, this pattern may be related to the fact that often ends of the viral coding regions tend to include various functional structures which naturally decrease the efficiency of the host CRISPR immune system (Rath et al., 2015); this corresponds to a weaker selection pressure for sequence similarity to the host.

4 Discussion

We suggest two major mechanisms that can explain the reported results (see **Supplementary Figure S9**): First, it is possible that the relation between long patterns in the viral coding sequences and viral fitness is related to the effect of these patterns on gene expression. Viral genomes include various types of motifs that are recognized by the host gene expression machinery; since the same (host) gene expression machinery processes both the viral and the host genes these motifs tend to appear both in the host and in the virus. Indeed, our analysis demonstrates that the long-subsequences that we find enrichment with sequence motifs (longer than singe codons) related to TFBS and RBPs.

Second, it is also possible that some of these patterns are related to the evolution of the virus for escaping the host immune system. It is important to emphasize that in our analysis the amino acid content of the viral genes was controlled for; thus, the reported signals cannot be, trivially, attributed only to the classical mechanisms, such as viral recognition by the host (*e.g.*, antibodies), as these mechanisms are traditionally believed to be based on interactions between proteins. However, it is plausible that they are related to alternative known and/or unknown immune mechanisms. One such relevant mechanism in bacteria is given by clustered regularly interspaced short palindromic repeats (CRISPR) (Krieg, 2002). This mechanism is based on creating fragments in the viral genome that are transcribed to short RNA molecules (crRNAs); these short RNA molecules match a certain region in the viral genome and 'guide' a protein complex (CAS-crRNA complex) that cuts the viral genome in this region and inactivates the virus. Since this mechanism is based on the recognition of short genomic sub-sequences that should appear in the virus/phage but not in the host, this may trigger evolution of the nucleotide composition of the virus/phage to be similar to the host. This may result in similar patterns of codons, and longer sequences that appear in the phage and the host, explaining especially high levels of AHRS-significant viruses in the bacteria reported here.

The fact that the enrichment with viral-host shared pattern is strongest in bacteria, in comparison to other viruses, may be related to various reasons: First, as discussed above, it may be related to viruses escaping the bacterial-specific immune mechanisms such as CRISPR. Second, it may be related to higher effective population size of bacteria and bacterial viruses, which is expected to contribute to higher selection efficiency (Kimura et al., 1963). Finally, this may be related to the fact that non-bacterial viruses tend to use more non-canonical gene expression regulatory mechanisms and codes.

Our analysis, demonstrate that the tendency to share sub-sequences with the host varies among proteins: Specifically, we have analyzed separately groups of proteins with different functions, found high enrichment for structural proteins (see **Figure 2**), and show that this result is not associated with the length of the virus ORFs. One explanation for that is related to the fact that these proteins tend to be more highly expressed and thus are under stronger selection for gene expression optimization, as is well known for non-viral genes; see for example, (dos Reis and Wernisch, 2009). In addition, our analysis shows that up to 15%-50% of the variance related to the shared host-virus sub-sequences can be explained by LDFs (*e.g.*, codon bias; see the Results). Among others, this correlation may be related to the fact that viruses that undergo stronger selection for LDFs *(e.g.*, due to larger effective population size or higher selection pressure) also tend to undergo stronger selection for shared long subsequences with the host in their coding region; for example, as explained above, both signals may contribute to improved expression levels.

It is important to emphasize, that similarly to viral adaptation to the host, silent features of the coding regions are expected to affect also related phenomenon, such as HGT. In this case a transferred gene is expected to be successfully expressed in a new host if its silent features are compatible (Medrano-Soto et al., 2004; Roller et al., 2013; Tuller, 2013, 2011; Tuller et al., 2011). Thus, although the host-homologous genes were excluded from our analysis, many of the results reported here may be generalized to the case of HGT. It is important to emphasize that a central HGT mechanism is transduction, the process in which bacterial DNA is moved from one bacterium to another by a bacteriophage (Soucy et al., 2015). Thus, the reported relations between 1) the host silent patterns and 2) the transferred gene silent patterns have much overlap: The fact that viral fitness is related to the similarity of its silent patterns to the host should directly improve its ability to transfer genes; it is also directly related to the fact that the silent aspects/codes in the transferred genes are more adapted to the new host since the virus undergoes evolution to be better adapted to the host.

Our results provide evidence of a complex, genomic level, evolutionary adaptation of viruses to their hosts and may have important implications for understanding viral evolution and for developing novel antiviral vaccines and therapeutic approaches. Various future direction and studies should be considered: First, the fitness and evolution of viruses can be tracked experimentally after decreasing and increasing their AVRS/AHVRS scores. Second, experimental and computational approaches for engineering viral coding regions for improving and decreasing their fitness based on the optimization of their AVRS/AHRS should be developed. Third, it will be interesting to perform further specific study related to the functionality of some of the virus-host repetitive sequences, or to the ways the host immune system may have been adapted to these silent/signals. This may require the deciphering of novel immune system pathways. Finally, it should be important to consider the pos-

sible effect of the non-trivial synonymous patterns reported here when developing models for viral molecular evolution; it may also be interesting and challenging to track the evolution of these patterns in viruses.

Acknowledgements

We thank Mr. Alon Diament, Dr. Hadas Zur, and Dr. Rachel Cohen-Kupiec for helpful discussions.

Funding

This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University and by the Minerva ARCHES award.

Conflict of Interest: none declared.

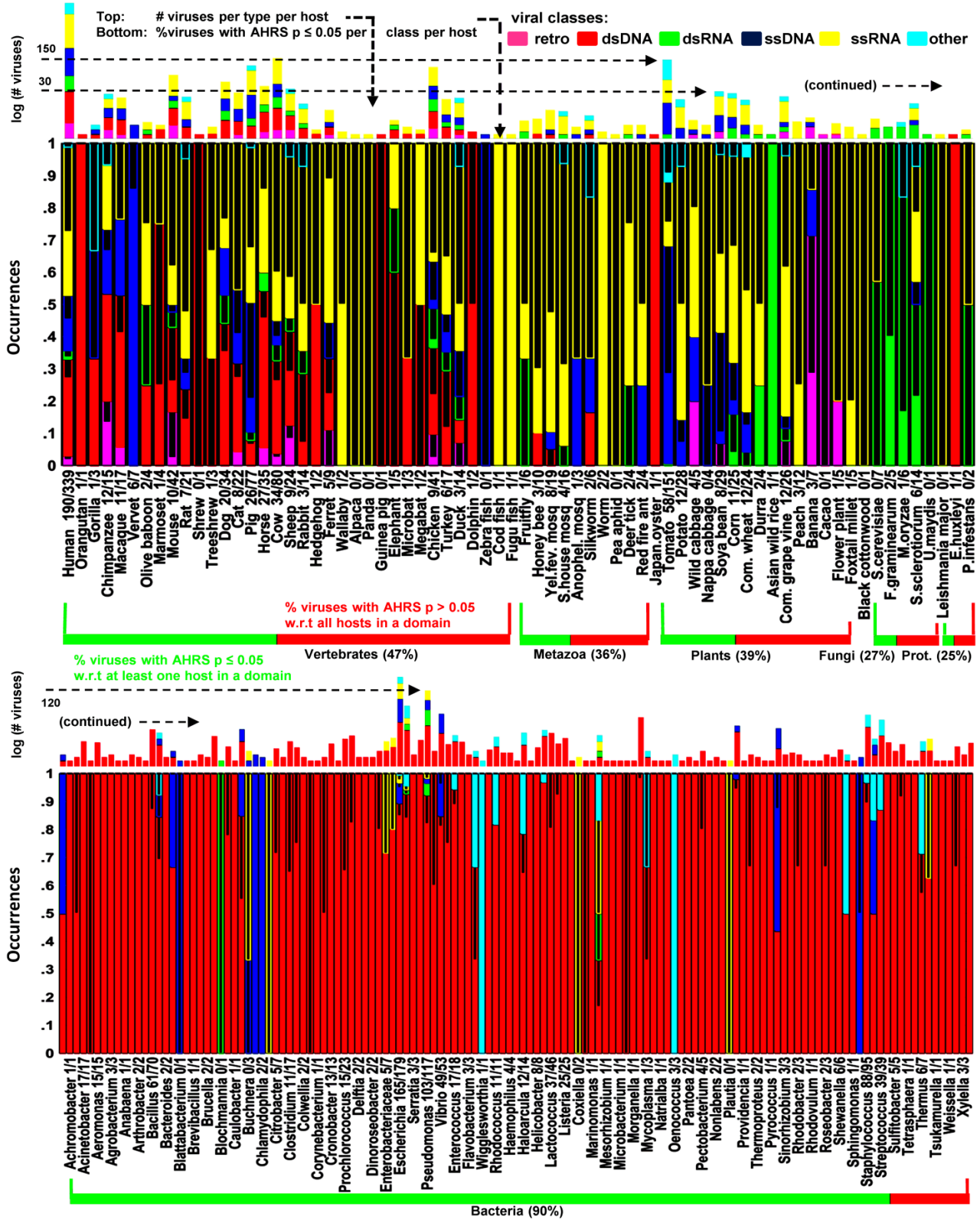

Figure 1: Selection for long host-repetitive patterns of silent functional information in viral coding regions. A summary of the analyzed hosts and viruses that undergo significant enrichment for mutually long sub-sequences. Each vertical bar corresponds to viruses infecting a specific host organism (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 AHRS-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 ($p\geq0.05$ w.r.t at least one randomization model). The numbers (*e.g.*, x/y) shown under each bar indicate the number of viruses (*e.g.*, x) that show significant enrichment out the total number of viruses checked (*e.g.*, y); 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.Horizontal bars visualizes the total percentage of AHRS-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 coding regions of at least one related host.

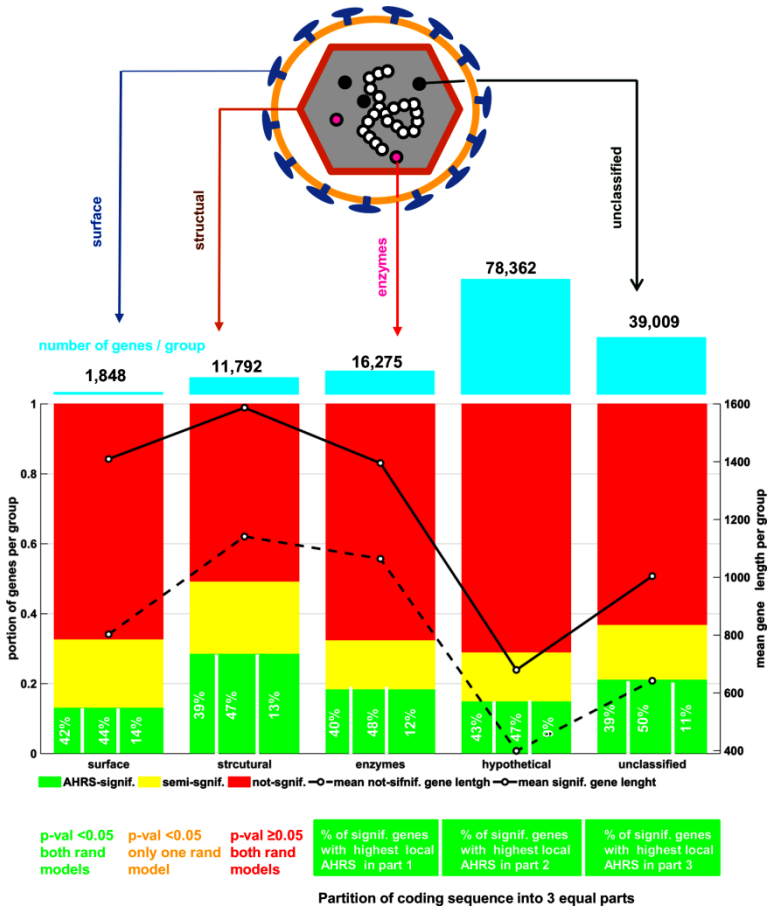

Figure 2: Selection for complex host-repetitive silent functional patterns depends on protein's function. The upper panel (in blue) represents the number of coding sequences within each functional group. The bars in the middle panel (green, yellow, and red, respectively) represent the percentage of significant (AHRS $p<0.05$ w.r.t both randomization models, green); semi-significant (AHRS $p<0.05$ w.r.t only one randomization models, green); non-significant (AHRS $p>0.05$ w.r.t. both randomization models). Black lines represent the mean length of significant (solid line) and non-significant (dotted-line) coding sequences in each group. We can see that those structural proteins are encoded by the highest portion of AHRS significant coding sequences. On the other hand, surface proteins have the smallest number of AHRS significant coding sequences. The enzymes and other proteins show an intermediate level of selection for long host-repetitive patterns. Each green bar (at the bottom) is divided into three parts, corresponding to the local AHRS analysis in the 5', middle, and 3' segments of a coding sequence. In each part the percentage of AHRS-significant genes with the highest local AHRS found in this part is indicated. We can see that for each gene group, most of the sequences used to have the highest local AHRS in the middle part; the percentage of genes with the highest local AHRS in the 3' part was found to be the smallest.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Bahir, I., Fromer, M., Prat, Y., Linial, M., 2009. VBahir, I., Fromer, M., Prat, Y., Linial, M., 2009. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Mol. Syst. Biol. 5, 311.iral adaptation to host: a proteome-based analysis of codon usage and amino acid pre. Mol. Syst. Biol. 5, 311.

Barrai, I., Salvatorelli, G., Mamolini, E., De Lorenzi, S., Carrieri, A., Rodriguez-Larralde, A., Scapoli, C., 2008. General preadaptation of viral infectors to their hosts. Intervirology 51, 101–11.

Barrai, I., Scapoli, C., Barale, R., Volinia, S., 1990. Oligonucleotide correlations between infector and host genomes hint at evolutionary relationships. Nucleic Acids Res. 18, 3021–5.

Brierley, I., 1995. Ribosomal frameshifting on viral RNAs. J. Gen. Virol. 76, 1885–1892.

Carbone, A., 2008. Codon bias is a major factor explaining phage evolution in translationally biased hosts. J. Mol. Evol. 66, 210–23.

Cardinale, D.J., Duffy, S., 2011. Single-stranded genomic architecture constrains optimal codon usage. Bacteriophage. 1(4):219-224.

Cheng, X., Wu, X., Wang, H., Sun, Y., Qian, Y., Luo, L., 2012. High codon adaptation in citrus tristeza virus to its citrus host. Virol. J. 9, 113.

Clyde, K., Harris, E., 2006. RNA secondary structure in the coding region of dengue virus type 2 directs translation start codon selection and is required for viral replication. J. Virol. 80, 2170–82.

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.

Cuevas, J.M., Domingo-Calap, P., Sanjuán, R., 2012. The fitness effects of synonymous mutations in DNA and RNA viruses. Mol. Biol. Evol. 29, 17–20.

Domingo, E., n.d. Virus as populations : composition, complexity, dynamics, and biological implications.

dos Reis, M. and Wernisch, L, 2009. Estimating translational selection in eukaryotic genomes. Mol. Biol. Evol. 26(2):451-461.

Firth, A.E., Brierley, I., 2012. Non-canonical translation in RNA viruses. J. Gen. Virol. 93, 1385–409.

Gale, M., Tan, S.-L., Katze, M.G., 2000. Translational Control of Viral Gene Expression in Eukaryotes. Microbiol. Mol. Biol. Rev. 64, 239–280.

Gibbs, A.J. (Adrian J.., Calisher, C.H., Garcia-Arenal, F., 2005. Molecular basis of virus evolution. Cambridge University Press.

Greenbaum, B.D., Levine, A.J., Bhanot, G., Rabadan, R., 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. 4, e1000079.

Groat-Carmona, A.M., Orozco, S., Friebe, P., Payne, A., Kramer, L., Harris, E., 2012. A novel coding-region RNA element modulates infectious dengue virus particle production in both mammalian and mosquito cells and regulates viral replication in Aedes aegypti mosquitoes. Virology 432, 511–526.

Gu, W., Zhou, T., Ma, J., Sun, X., Lu, Z., 2004. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Res. 101, 155–161.

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, Mol. Cell, 38, 576-589.

Holmes, E.C., Drummond, A.J., 2007. The evolutionary genetics of viral emergence.

Curr. Top. Microbiol. Immunol. 315, 51–66.

Jackson, R.J., 2005. Alternative mechanisms of initiating translation of mammalian mRNAs. Biochem. Soc. Trans. 33, 1231–41.

Jenkins, G.M., Pagel, M., Gould, E.A., de A Zanotto, P.M., Holmes, E.C., 2001. Evolution of base composition and codon usage bias in the genus Flavivirus. J. Mol. Evol. 52, 383–90.

Kerr, J.R., Boschetti, N., 2006. Short regions of sequence identity between the genomes of human and rodent parvoviruses and their respective hosts occur within host genes for the cytoskeleton, cell adhesion and Wnt signalling. J. Gen. Virol. 87, 3567–75.

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.

Kieft, J.S., 2008. Viral IRES RNA structures and ribosome interactions. Trends Biochem. Sci. 33, 274–83.

Kimura, M., Maruyama, T., Crow, J.F., 1963. The Mutation Load in Small Populations. Genetics. 48(10), 1303–1312.

Krieg, A.M., 2002. CpG motifs in bacterial DNA and their immune effects. Annu. Rev. Immunol. 20, 709–760.

Kunec, D., Osterrieder, N., 2016. Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias. Cell Rep. 14, 55–67.

Lobo, F.P., Mota, B.E.F., Pena, S.D.J., Azevedo, V., Macedo, A.M., Tauch, A., Machado, C.R., Franco, G.R., 2009. Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts. PLoS One 4, e6282.

López-Lastra, M., Ramdohr, P., Letelier, A., Vallejos, M., Vera-Otarola, J., Valiente-Echeverría, F., 2010. Translation initiation of viral mRNAs. Rev. Med. Virol. 20, 177–195.

Lucks, J.B., Nelson, D.R., Kudla, G.R., Plotkin, J.B., 2008. Genome landscapes and bacteriophage codon usage. PLoS Comput. Biol. 4, e1000001.

Manber, U., Myers, G., 1993. Suffix Arrays: A New Method for On-Line String Searches. SIAM J. Comput. 22, 935–948.

Medrano-Soto, A., Moreno-Hagelsieb, G., Vinuesa, P., Christen, J.A., Collado-Vides, J., 2004. Successful Lateral Transfer Requires Codon Usage Compatibility Between Foreign Genes and Recipient Genomes. Mol. Biol. Evol. 21, 1884–1894.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., Ogata, H., 2016. Linking Virus Genomes with Host Taxonomy. Viruses 8, 66.

Mueller, S., Papamichail, D., Coleman, J.R., Skiena, S., Wimmer, E., 2006. Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. J. Virol. 80, 9687–96.

Paz, I., Kosti, I., Ares, M., Cline, M., Mandel-Gutfreund, Y., 2014. RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res. 42, W361-W367.

Pride, D., Wassenaar, T., Ghose, C., Blaser, M., 2006. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 7, 8.

Rath, D., Amlinger, L., Rath, A., Lundgren, M., 2015. The CRISPR-Cas immune system: Biology, mechanisms and applications. Biochimie 117, 119–128.

Roller, M., Lucic, V., Nagy, I., Perica, T., Vlahovicek, K., 2013. Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic Acids Res. 41, 8842–8852.

Sau, K., Gupta, S.K., Sau, S., Mandal, S.C., Ghosh, T.C., 2007. Studies on synonymous codon and amino acid usage biases in the broad-host range bacteriophage KVP40. J. Microbiol. 45, 58–63.

Sau, K., Sau, S., Mandal, S.C., Ghosh, T.C., 2005. Factors influencing the synonymous codon and amino acid usage bias in AT-rich Pseudomonas aeruginosa phage PhiKZ. Acta Biochim. Biophys. Sin. (Shanghai). 37, 625–33.

Shackelton, L.A., Parrish, C.R., Holmes, E.C., 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J. Mol. Evol. 62, 551–63.

Soucy, S.M., Huang, J., Gogarten, J.P., 2015. Horizontal gene transfer: building the web of life. Nat. Rev. Genet. 16, 472–482.

Su, M.-W., Lin, H.-M., Yuan, H.S., Chu, W.-C., 2009. Categorizing host-dependent RNA viruses by principal component analysis of their codon usage preferences. J. Comput. Biol. 16, 1539–47.

Thurner, C., Witwer, C., Hofacker, I.L., Stadler, P.F., 2004. Conserved RNA secondary structures in Flaviviridae genomes. J. Gen. Virol. 85, 1113–24.

Tuller, T., 2011. Codon bias, tRNA pools and horizontal gene transfer. Mob. Genet. Elements 1, 75–77.

Tuller, T., 2013. The Effect of Codon Usage on the Success of Horizontal Gene Transfer. In: Lateral Gene Transfer in Evolution. Springer New York, New York, NY, pp. 147–158.

Tuller, T., Girshovich, Y., Sella, Y., Kreimer, A., Freilich, S., Kupiec, M., Gophna, U., Ruppin, E., 2011. Association between translation efficiency and horizontal gene transfer within microbial communities. Nucleic Acids Res. 39, 4743–55.

Ulitsky, I., Burstein, D., Tuller, T., Chor, B., 2006. The average common substring approach to phylogenomic reconstruction. J. Comput. Biol. 13, 336–50.

van Hemert, F.J., Berkhout, B., Lukashov, V. V., 2007. Host-related nucleotide composition and codon usage as driving forces in the recent evolution of the Astroviridae. Virology 361, 447–454.

Zafrir, Z., Tuller, T., 2017. Unsupervised detection of regulatory gene expression information in different genomic regions enables gene expression ranking. BMC Bioinformatics 18, 77.

Zhao, S., Zhang, Q., Liu, X., Wang, X., Zhang, H., Wu, Y., Jiang, F., 2008. Analysis of synonymous codon usage in 11 Human Bocavirus isolates. Biosystems 92, 207–214.

Ziv, J., Lempel, A., 1977. A universal algorithm for sequential data compression. IEEE Trans. Inf. Theory 23, 337–343.

Zur, H., Tuller, T., 2015. Exploiting hidden information interleaved in the redundancy of the genetic code without prior knowledge. Bioinformatics 31, 1161–8.

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

Eli Goz[1,2,&], Zohar Zafrir[1,2,&], Tamir Tuller[1,2,3,*]

[1]Department of Biomedical Engineering, Tel-Aviv University, Ramat Aviv, Israel. [2]SynVaccineLtd .Ramat Hachayal, Tel Aviv, Israel.[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. 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)**.

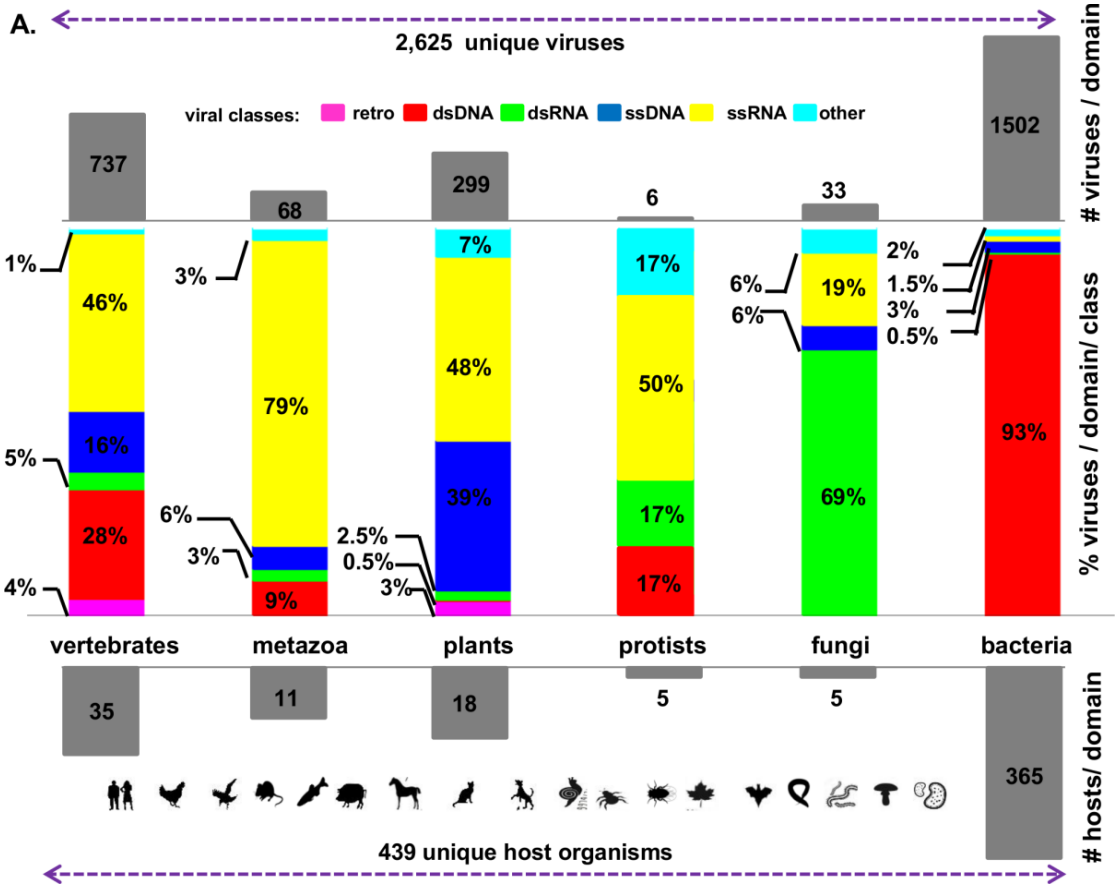

Figure S1: Viruses-hosts dataset summary. We analyzed 2,625 unique viruses belonging to different Baltimore classes: reverse-transcribing (retro), double-stranded DNA (dsDNA), double-stranded RNA (dsRNA), single-stranded DNA (ssDNA), single stranded RNA (ssRNA, positive and negative sense) and other (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:

$$\log(|G|)\,/-\mathrm{E}_{\mathrm{M_S}}\log\left(\mathrm{M_G}\right)$$

$$-\mathrm{E}_{\mathrm{M_S}}\log(\mathrm{M_G}) = \lim_{\mathrm{n}\to\infty}\left(\sum\nolimits_{\mathrm{x^n}} \mathrm{M_s}(\mathrm{x^n})\log\left(\frac{1}{\mathrm{M_G}(\mathrm{x^n})}\right)\right)$$

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

$$\log(|G|)\,/\mathrm{H}(\mathrm{M_S})$$

where $\mathrm{H}(\mathrm{M_S})$ is the entropy of $\mathrm{M_S}$ and it is known that H(MS) is smaller than $-\mathrm{E}_{\mathrm{M_S}}\log\left(\mathrm{M_G}\right)$ (second equation) for Ms ≠MG. 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.

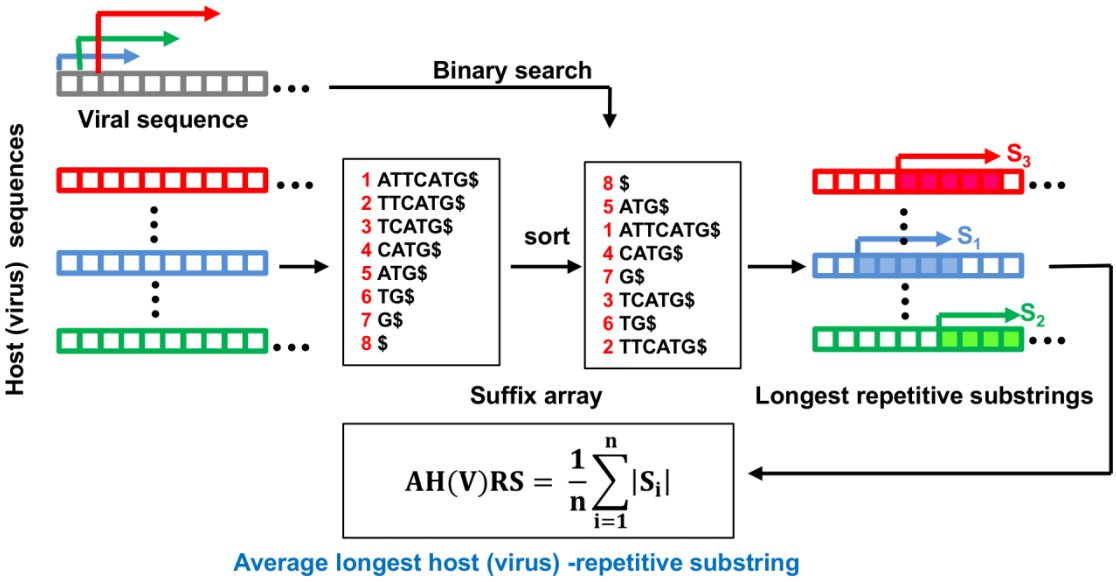

Figure S2: **Average host (virus) – repetitive scores**

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.

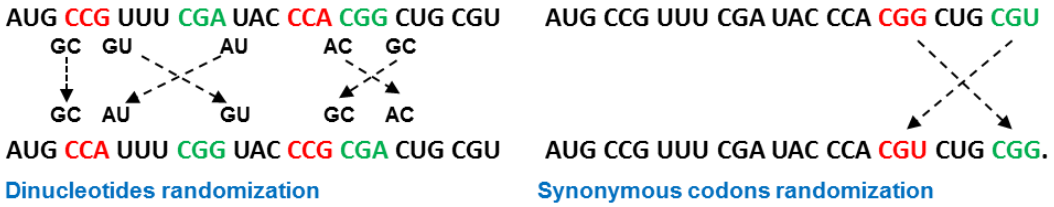

Figure S3: Randomization models

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:

$$\text{p-value} = \frac{|\,\text{rand}_i\text{: rand}_i \geq \text{x})\,|}{\text{N+1}}$$

$$\text{z-score} = \frac{\text{x- mean}_i[\text{rand}_i]}{\text{std}_i[\text{rand}_i]}$$

1.5 Analysis of low-dimensional features

The low-dimensional features were computes 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:

$$n = \sum_{i}^{d} x_i$$

The frequency of each codon is therefore:

$$p_i = x_i/n$$

The ENC for a specific AA is:

$\widehat{N_e} = 1/\hat{F}$, where $\hat{F} = \sum_i^d p_i^2$

ENC for the group of AAs with degeneracy *d*:

$N = 1/\overline{\widehat{F_d}}$, where $\overline{\widehat{F_d}} = \frac{1}{|A_d|}\sum_{i\in A_d} \hat{F}_i$

(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:

$\widehat{N_e} = 2 + \frac{9}{\widehat{F_2}} + \frac{1}{\widehat{F_3}} + \frac{5}{\widehat{F_4}} + \frac{3}{\widehat{F_6}}$,

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:

$$CPS = \log\left(\frac{F(AB)}{\frac{F(A)\times F(B)}{F(X)\times F(Y)}\times F(XY)}\right),$$

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 than defined as an avenge 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):

$$Similarity\ Score = \frac{1}{Motif\ Length} \sum_{i}^{Motif\ Length} -\frac{(Observed_i - Expect_i)}{Expect_i}$$

$$Observed_i = \sum_{j}^{A,C,G,T} -\left(freq_1^{ij} - freq_2^{ij}\right)^2$$

$$Expect_i = \sum_{j}^{A,C,G,T} \sum_{k}^{A,C,G,T} -\frac{\left(freq_1^{ij} - freq_2^{ik}\right)^2}{4}$$

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 S4VII**I). 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**).

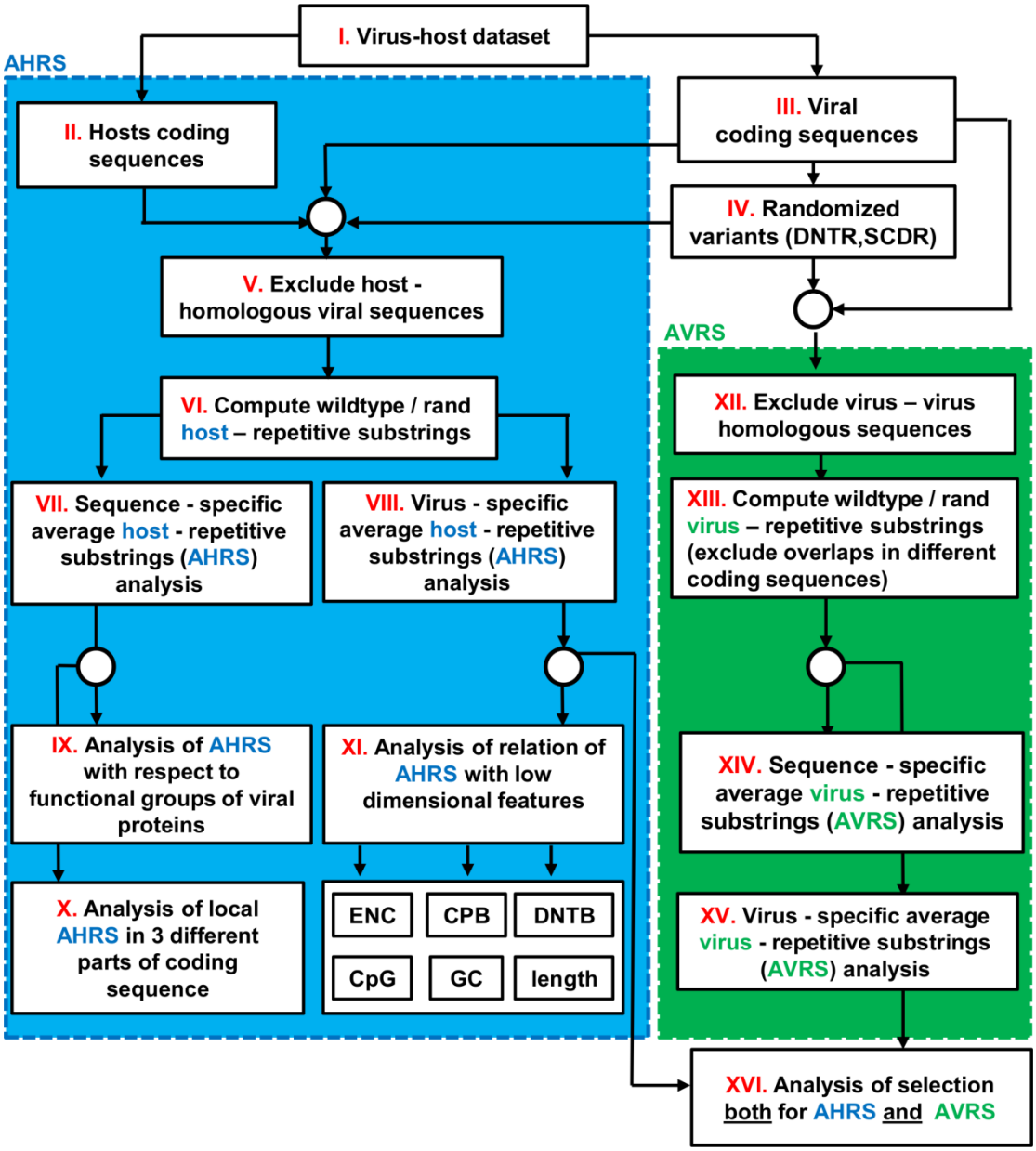

Figure S4: Flow diagram of the study.

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.

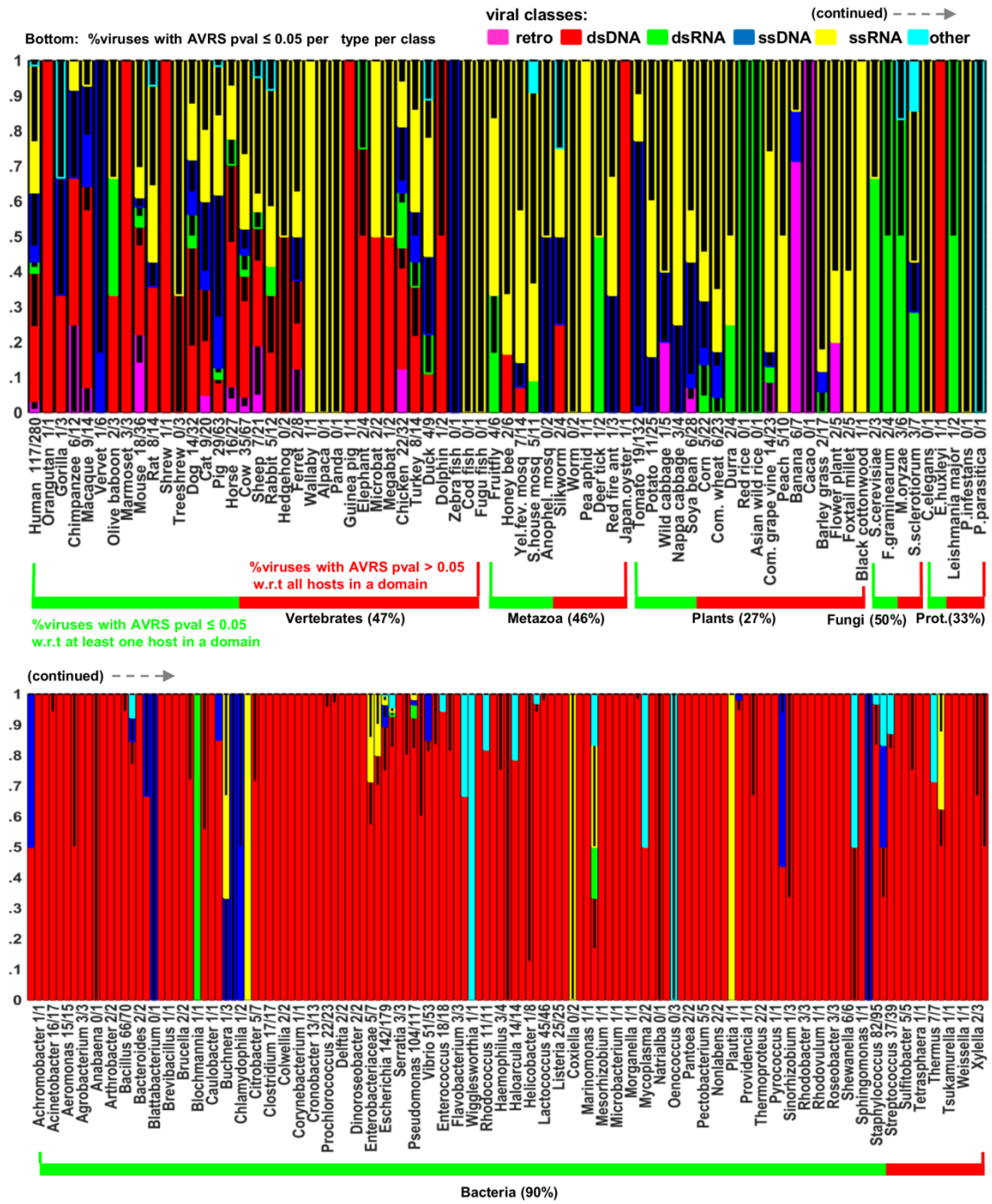

Figure S5: Selection for long virus–repetitive patterns of silent functional information in viral coding regions. Each vertical bar corresponds to viruses infecting a specific host organism (in bacteria – a specific genus) and is partitioned into class specific segments; every segment corresponds to 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 ($p\geq0.05$ 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, meatzoa, 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.

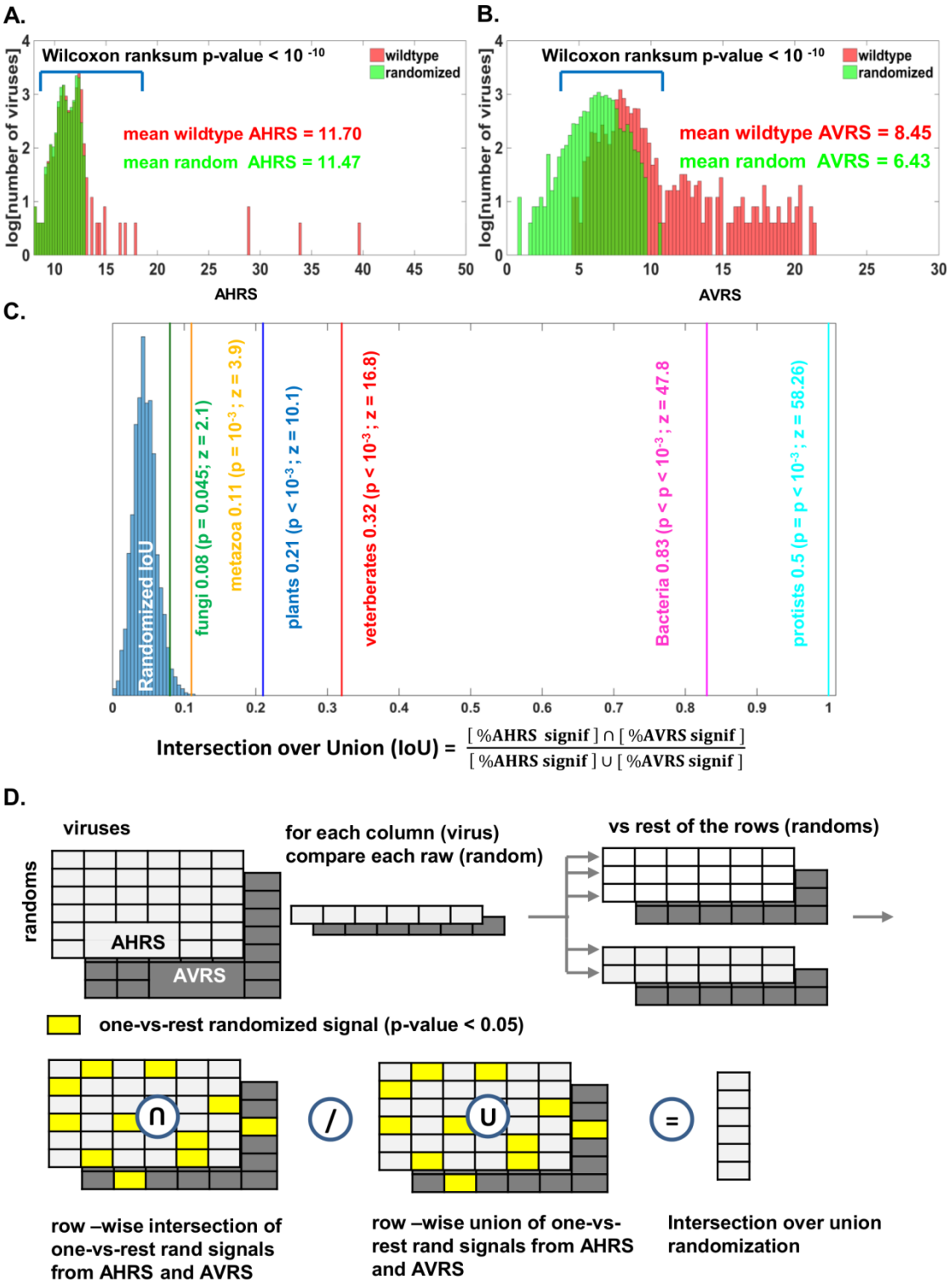

Figure S6: A. Distribution of AHRS values for wildtype and randomized viral coding sequences. The average wildtype AHRS value (11.70) is higher than the average randomized (11.47) and this relation is statistically significant (Wilcoxon rank-sum right tail $p<10^{-10}$). **B. Distribution of AVRS values 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

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 tow 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 than, 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**).

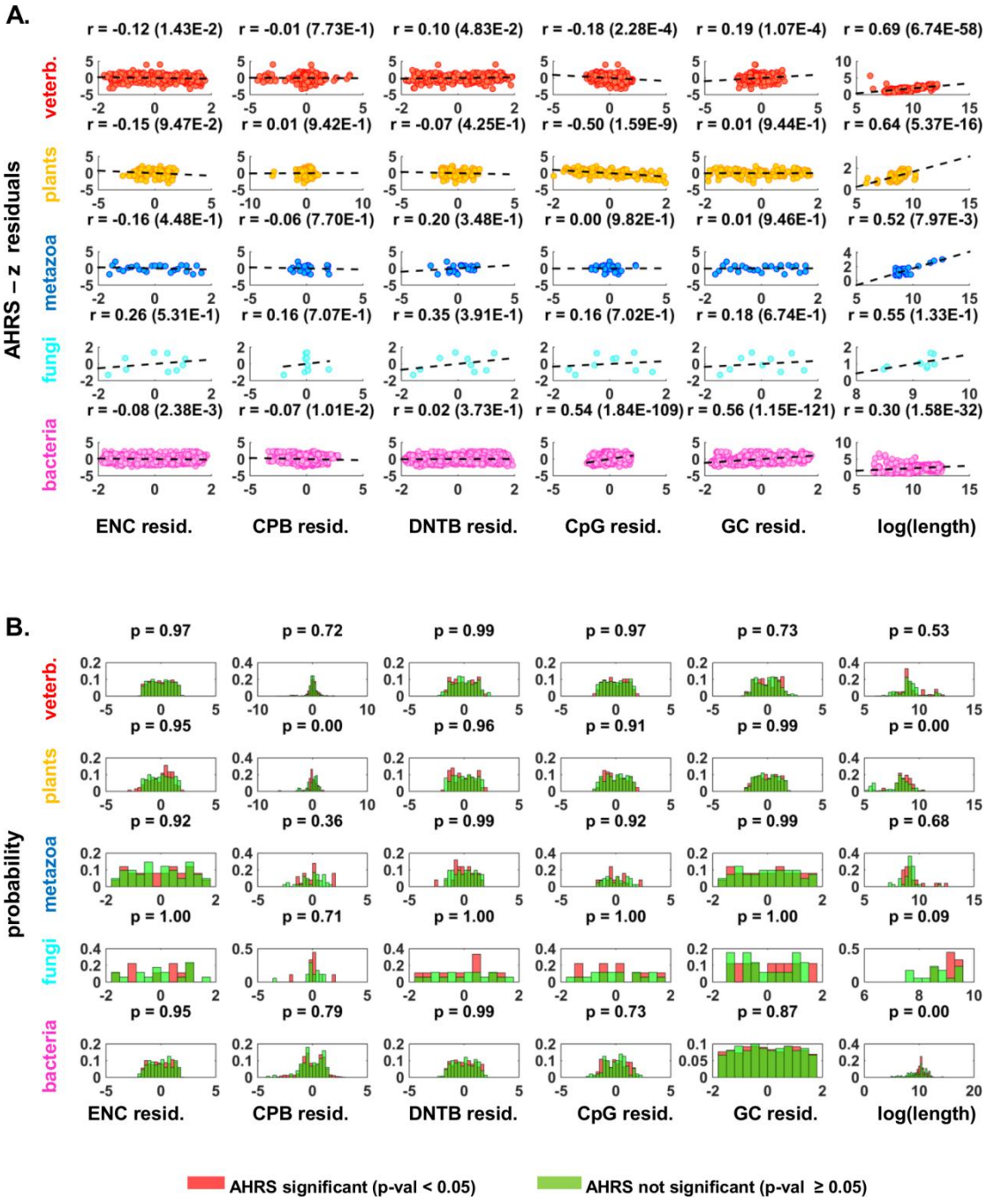

Figure S7: Relation between low dimensional features and AHRS z-scores. LDF values and corresponding AHRS-z scores (AHRS-z) may be significantly correlated (either positively or negatively) due to a positive correlation of both of them with the genome length. Therefore, to control for the viral genome 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

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 ($p<4.58\cdot10^{-7}$); 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**.

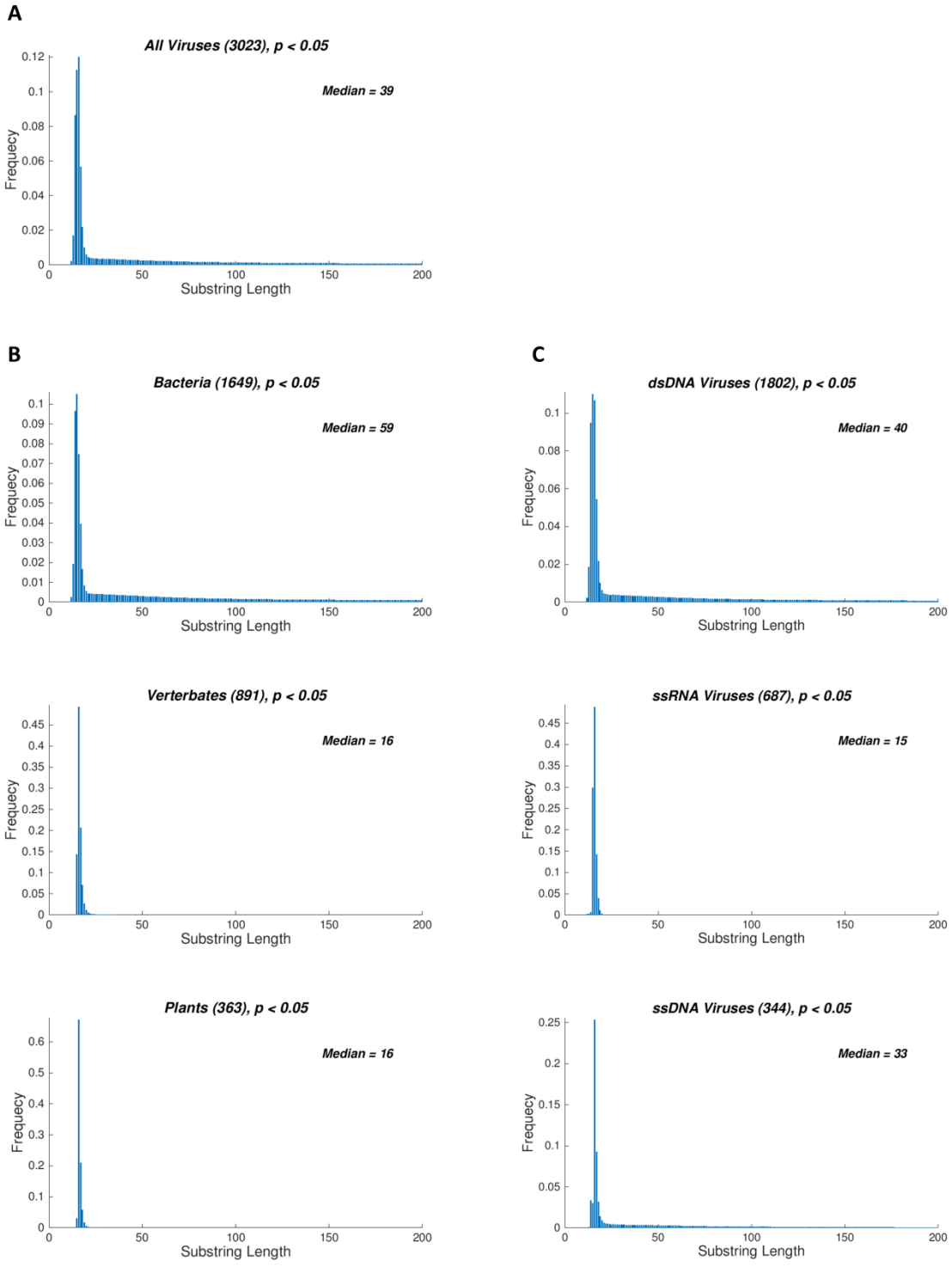

Figure S8: Analysis of length distribution of significantly enriched substrings. A. All Viruses. **B.** Division according to host type. **C.** Division according to virus type.

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

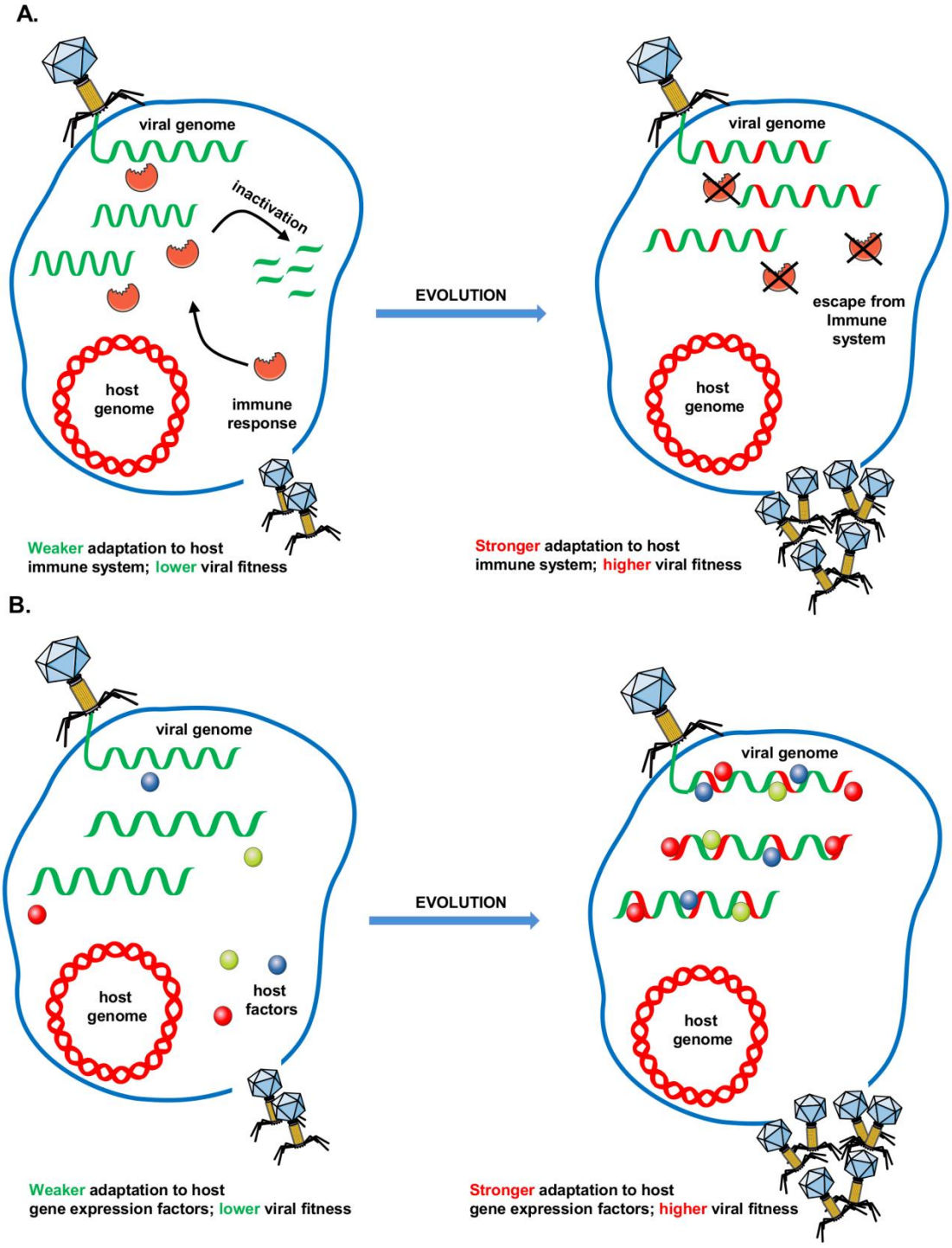

Figure S9: Two suggested non-mutually exclusive hypotheses related to the observed tendency of viruses to include long sub-sequences/codes in their coding regions that appear also in the host. A. These codes enable efficient immune system avoidance; specifically they may enable escaping the CRISPR system. **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 genomes collection [WWW Document], n.d. URL http://www.ensembl.org/info/about/species.html#eg

Ensembl Metazoan 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., Shimizu, 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. RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res. 42, W361-W367.

SPARCS webpagge [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.