Bioinformatics, 34(19), 2018, 3241–3248
doi: 10.1093/bioinformatics/bty351
Advance Access Publication Date: 30 April 2018
Discovery Note

Genome analysis

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,†] and Tamir Tuller[1,2,3,*]

[1]Department of Biomedical Engineering, Tel Aviv University, Tel Aviv 6997801, Israel, [2]SynVaccineLtd, Ramat Hachayal, Tel Aviv 6997801, Israel and [3]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv 6997801, Israel

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on September 7, 2017; revised on March 1, 2018; editorial decision on April 25, 2018; accepted on April 27, 2018

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 2625 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

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, 2005; 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 does not 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) and 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) and Shackelton *et al.* (2006) it was shown that CpG pairs are underrepresented 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.*, 2005, 2007; 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 10 nt 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 2625 unique viruses of all classes and 439 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 Materials and methods

In this section we briefly summarize the most important rationale of our methodology. The details appear in Supplementary Sections S1.1–S1.7.

2.1 Data preparation

The associations of viruses to their host organisms were retrieved from the GenomeNet Virus-Host Database (virus-host DB) (Mihara *et al.*, 2016). In total we collected 2625 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 S1.1).

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 a single 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 S1.2): (i) Build a suffix array (Manber and Myers, 1993)-this can be done in $O(|H(V)|)$ (Gusfield, 1997); (ii) For each position i in a viral coding sequence S, use the suffix array from (i) 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.); (iii) 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 cannot 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: (i) Dinucleotide Randomization that preserves both the amino acids order and content, and the frequency distribution of 16 possible pairs of adjacent nucleotides (dinucleotides); (ii) Synonymous Codon Randomization that preserves the amino acids order and content, mono-nucleotide composition and the codon usage bias (see also Supplementary Section S1.3).

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 S1.4).

3 Results

3.1 Overview of the analysis

The general stages of our study are as follows (see more details in Supplementary Section S2.1 and Supplementary Fig. 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 1000 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 empirical *P*-values and Z-scores with respect to both randomization models were computed for each viral coding region separately. In addition, 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 S2.4 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 ($P < 4.58 \times 10^{-7}$). 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 S1.5 and S2.3 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

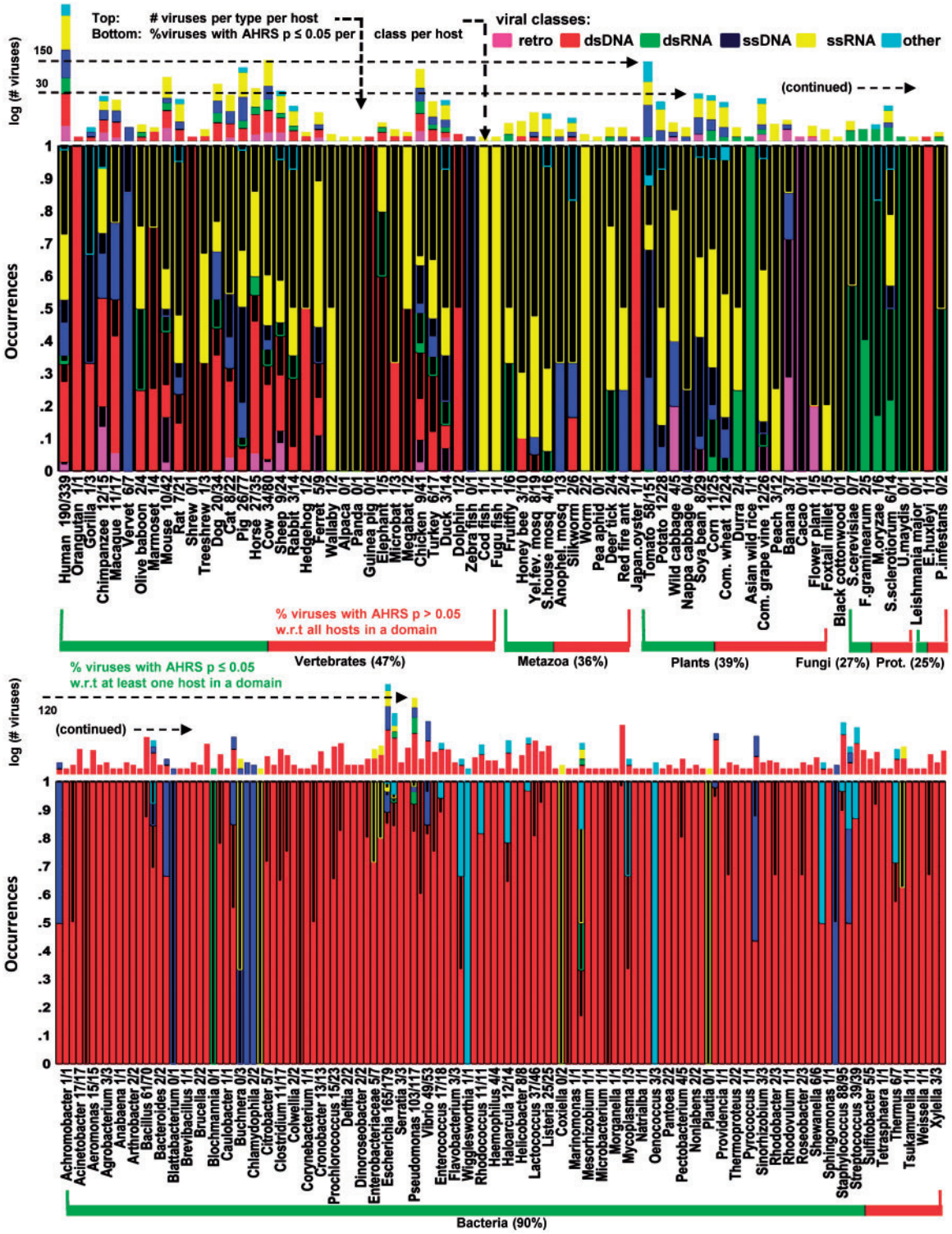

Fig. 1. Selection for long host-repetitive patterns of silent functional information in viral coding regions. A summary of the analyzed hosts and viruses that undergo significant enrichment for mutually long sub-sequences. Each vertical bar corresponds to viruses infecting a specific host organism (in 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 ($P \geq 0.05$ 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 the 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

coding regions of at least one related host (Fig. 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, metazoa 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 S2.2 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.*, 2010) that appear in human viruses more than expected by the our null model (see Supplementary Section S1.7). Next, we compared these motifs against known information of TFBS and RBPs, taken from the JASPAR (Khan *et al.*, 2018) 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 found that generally these viral genomes tend to include more TF and RBP binding sites than expected from a Null model ($P < 0.04$); see more details in Supplementary Section S1.8 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 finding 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 five mutually exclusive functional groups (see also Supplementary Section S1.6): surface genes, structural genes, enzymes, hypothetical (putative proteins) and unclassified (accessory or regulatory

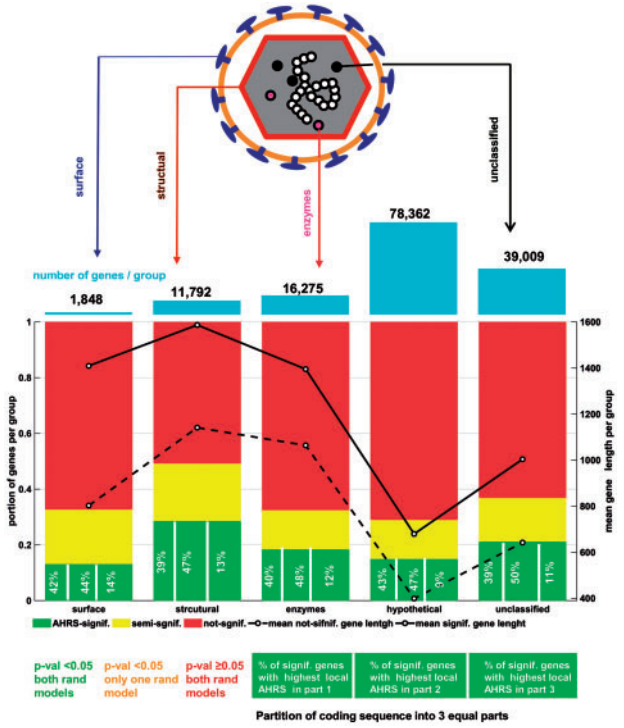

Fig. 2. Selection for complex host-repetitive silent functional patterns depends on protein's function. The upper panel (in blue) represents the number of coding sequences within each functional group. The bars in the middle panel (green, yellow and red, respectively) represent the percentage of significant (AHRS $P < 0.05$ w.r.t both randomization model, 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

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 $P < 1 \times 10^{-16}$). 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 cannot be attributed only to sequence lengths, the analyzed viral coding regions were divided into four 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 contain the lowest number. Therefore, our conclusions cannot be attributed only to the lengths of the coding regions (see details in Supplementary Section S2.6). 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 three 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 Fig. 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 found are enriched 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 from 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 the 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 in 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 demonstrates that the tendency to share subsequences with the host varies among proteins. Specifically, we have analyzed separately groups of proteins with different functions, found high enrichment for structural proteins (see Fig. 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 (i) the host silent patterns and (ii) 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 possible 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.

References

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Bahir,I. *et al*. (2009) Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol. Syst. Biol*, **5**, 311.

Barrai,I. *et al*. (2008) General preadaptation of viral infectors to their hosts. *Intervirology*, **51**, 101–111.

Barrai,I. *et al*. (1990) Oligonucleotide correlations between infector and host genomes hint at evolutionary relationships. *Nucleic Acids Res*., **18**, 3021–3025.

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**, 210–223.

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

Cheng,X. *et al*. (2012) High codon adaptation in citrus tristeza virus to its citrus host. *Virol. J*., **9**, 113.

Clyde,K. and 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**, 2170–2182.

Coleman,J.R. *et al*. (2008) Virus attenuation by genome-scale changes in codon pair bias. *Science*, **320**, 1784–1787.

Cuevas,J.M. *et al*. (2012) The fitness effects of synonymous mutations in DNA and RNA viruses. *Mol. Biol. Evol*., **29**, 17–20.

Domingo,E. (2005) *Virus as populations: composition, complexity, dynamics, and biological implications*. Academic Press, USA.

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

Firth,A.E. and Brierley,I. (2012) Non-canonical translation in RNA viruses. *J. Gen. Virol*., **93**, 1385–1409.

Gale,M. *et al*. (2000) Translational control of viral gene expression in eukaryotes. *Microbiol. Mol. Biol. Rev*., **64**, 239–280.

Gibbs,A.J. *et al*. (2005) *Molecular Basis of Virus Evolution*. Cambridge University Press Cambridge, UK.

Greenbaum,B.D. *et al*. (2008) Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog*., **4**, e1000079.

Groat-Carmona,A.M. *et al*. (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. *et al*. (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 Cambridge, UK.

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. and 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**, 1231–1241.

Jenkins,G.M. *et al*. (2001) Evolution of base composition and codon usage bias in the genus Flavivirus. *J. Mol. Evol*., **52**, 383–390.

Kerr,J.R. and 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**, 3567–3575.

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**, 274–283.

Kimura,M. *et al*. (1963) The mutation load in small populations. *Genetics*, **48**, 1303–1312.

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

Kunec,D. and Osterrieder,N. (2016) Codon pair bias is a direct consequence of dinucleotide bias. *Cell Rep*., **14**, 55–67.

Lobo,F.P. *et al*. (2009) Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts. *PLoS One*, **4**, e6282.

López-Lastra,M. *et al*. (2010) Translation initiation of viral mRNAs. *Rev. Med. Virol*., **20**, 177–195.

Lucks,J.B. *et al*. (2008) Genome landscapes and bacteriophage codon usage. *PLoS Comput. Biol*., **4**, e1000001.

Manber,U. and Myers,G. (1993) Suffix arrays: a new method for on-line string searches. *SIAM J. Comput*., **22**, 935–948.

Medrano-Soto,A. *et al*. (2004) Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. *Mol. Biol. Evol*., **21**, 1884–1894.

Mihara,T. *et al*. (2016) Linking virus genomes with host taxonomy. *Viruses*, **8**, 66.

Mueller,S. *et al*. (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**, 9687–9696.

Paz,I. *et al*. (2014) RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res*., **42**, W361–W367.

Pride,D. *et al*. (2006) Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics*, **7**, 8.

Rath,D. *et al*. (2015) The CRISPR-Cas immune system: biology, mechanisms and applications. *Biochimie*, **117**, 119–128.

Roller,M. *et al*. (2013) Environmental shaping of codon usage and functional adaptation across microbial communities. *Nucleic Acids Res*., **41**, 8842–8852.

Sau,K. *et al*. (2007) Studies on synonymous codon and amino acid usage biases in the broad-host range bacteriophage KVP40. *J. Microbiol*., **45**, 58–63.

Sau,K. *et al*. (2005) Factors influencing the synonymous codon and amino acid usage bias in AT-rich Pseudomonas aeruginosa phage PhiKZ. *Acta Biochim. Biophys. Sin. (Shanghai)*, **37**, 625–633.

Shackelton,L.A. *et al*. (2006) Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *J. Mol. Evol*., **62**, 551–563.

Soucy,S.M. *et al*. (2015) Horizontal gene transfer: building the web of life. *Nat. Rev. Genet*., **16**, 472–482.

Su,M.-W. *et al*. (2009) Categorizing host-dependent RNA viruses by principal component analysis of their codon usage preferences. *J. Comput. Biol*., **16**, 1539–1547.

Thurner,C. *et al*. (2004) Conserved RNA secondary structures in Flaviviridae genomes. *J. Gen. Virol*., **85**, 1113–1124.

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: Gophna,U. (ed.) *Lateral Gene Transfer in Evolution*. Springer, New York, NY, pp. 147–158.

Tuller,T. *et al*. (2011) Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Res*., **39**, 4743–4755.

Ulitsky,I. *et al*. (2006) The average common substring approach to phylogenomic reconstruction. *J. Comput. Biol*., **13**, 336–350.

van Hemert,F.J. *et al*. (2007) Host-related nucleotide composition and codon usage as driving forces in the recent evolution of the Astroviridae. *Virology*, **361**, 447–454.

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

Zhao,S. *et al*. (2008) Analysis of synonymous codon usage in 11 Human Bocavirus isolates. *Biosystems*, **92**, 207–214.

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

Zur,H. and Tuller,T. (2015) Exploiting hidden information interleaved in the redundancy of the genetic code without prior knowledge. *Bioinformatics*, **31**, 1161–1168.