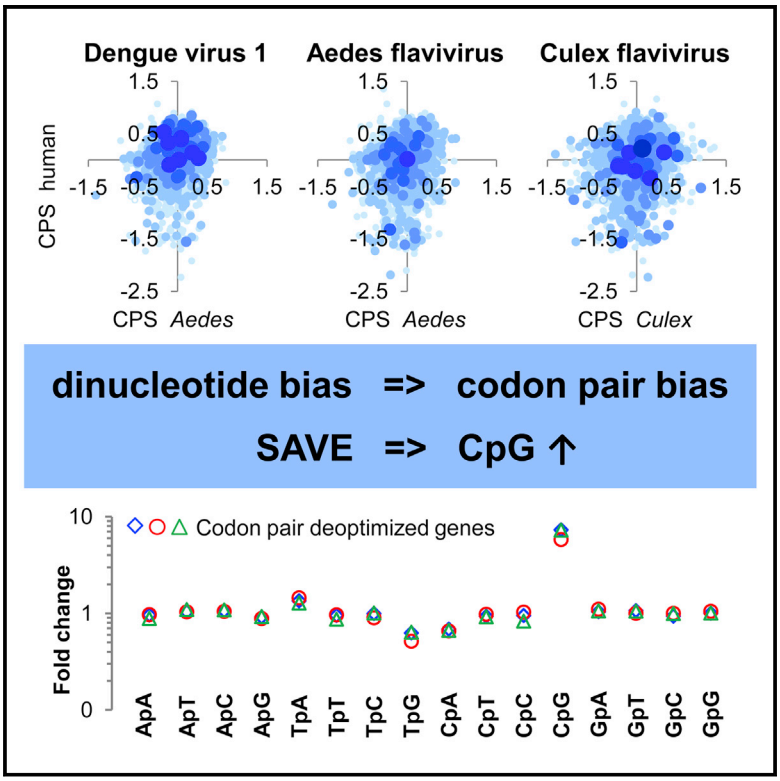
Article

Cell Reports

Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias

Graphical Abstract



Highlights

- Encoding in human viruses is only marginally influenced by host codon pair bias
- Arthropod-specific viruses do not share the codon pair bias of their hosts
- Codon pair bias is essentially a direct consequence of dinucleotide bias
- Attenuation by SAVE is achieved by an increase in CpG dinucleotides in recoded genes

Authors

Dusan Kunec, Nikolaus Osterrieder

Correspondence

dusan.kunec@fu-berlin.de (D.K.),
no.34@fu-berlin.de (N.O.)

In Brief

Kunec and Osterrieder demonstrate that the encoding of viral proteins is not influenced by codon pair preferences of their host but that it can be influenced by host dinucleotide bias. Codon pair bias is primarily a consequence of dinucleotide bias. Attenuation by codon pair deoptimization works through an increase in CpG dinucleotides in recoded genes.

Kunec & Osterrieder, 2016, Cell Reports *14*, 1–13
January 5, 2016 ©2016 The Authors
http://dx.doi.org/10.1016/j.celrep.2015.12.011

CellPress

Please cite this article in press as: Kunec and Osterrieder, Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias, Cell Reports (2016), http://dx.doi.org/10.1016/j.celrep.2015.12.011

Cell Reports
Article

OPEN ACCESS
CellPress

Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias

Dusan Kunec[1,*] and Nikolaus Osterrieder[1,*]
[1]Institut für Virologie, Zentrum für Infektionsmedizin, Freie Universität Berlin, Robert-von-Ostertag-Straße 7–13, 14163 Berlin, Germany
*Correspondence: dusan.kunec@fu-berlin.de (D.K.), no.34@fu-berlin.de (N.O.)
http://dx.doi.org/10.1016/j.celrep.2015.12.011
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

SUMMARY

Codon pair bias is a remarkably stable characteristic of a species. Although functionally uncharacterized, robust virus attenuation was achieved by recoding of viral proteins using underrepresented codon pairs. Because viruses replicate exclusively inside living cells, we posited that their codon pair preferences reflect those of their host(s). Analysis of many human viruses showed, however, that the encoding of viruses is influenced only marginally by host codon pair preferences. Furthermore, examination of codon pair preferences of vertebrate, insect, and arthropod-borne viruses revealed that the latter do not utilize codon pairs overrepresented in arthropods more frequently than other viruses. We found, however, that codon pair bias is a direct consequence of dinucleotide bias. We conclude that codon pair bias does not play a major role in the encoding of viral proteins and that virus attenuation by codon pair deoptimization has the same molecular underpinnings as attenuation based on an increase in CpG/TpA dinucleotides.

INTRODUCTION

Attenuation by codon pair deoptimization has emerged recently as a strategy for rapid and highly efficacious attenuation of various small RNA viruses (Coleman et al., 2008; Le Nouën et al., 2014; Mueller et al., 2010; Shen et al., 2015; Wang et al., 2015; Yang et al., 2013). The strategy, also known as synthetic attenuated virus engineering (SAVE), eliminates many of the drawbacks of traditional vaccine development and has resulted in the generation of superior experimental live virus vaccines (Coleman et al., 2008; Le Nouën et al., 2014; Mueller et al., 2010; Shen et al., 2015; Wang et al., 2015; Yang et al., 2013). Attenuation by SAVE is based on large-scale recoding of viral genes while precisely preserving the amino acid sequences of the encoded proteins.

The actual encoding of amino acids is biased, and some codons are used more often than others, a phenomenon known as codon bias. Similarly, but independently of codon bias, juxtaposition of codons in open reading frames (ORFs) appears to be not random either (Gutman and Hatfield, 1989). Some codon pairs are found in ORFs significantly more or less frequently than would be expected based on the overall frequencies of two codons that form a particular codon pair or bicodon (Coleman et al., 2008; Gutman and Hatfield, 1989; Mueller et al., 2006). These preferences are typically referred to as codon pair preference or codon pair bias. Codon pair preference has been found in every species studied and can be radically dissimilar between phylogenetically distant species (Moura et al., 2005; Mueller et al., 2010). Its existence has been known for many years, but its biological significance and the forces that shape this bias are only poorly understood (Moura et al., 2005).

The attenuation by SAVE is achieved through the reshuffling of existing synonymous codons in a coding sequence. The goal is to increase the number of codon pairs that are underrepresented in the protein coding sequences of the host because these are implicated in creating unfavorable conditions for protein production, processing, or folding (Coleman et al., 2008; Shen et al., 2015). This hypothesis was never thoroughly tested experimentally. Following the logic of the hypothesis, the procedure directly causes a reduction of the reproductive fitness of the virus and attenuation (Coleman et al., 2008).

Because viruses replicate exclusively inside of living cells and depend on the protein synthesis and chaperone machineries of the host, we speculated that the primary structure of viral genes might be shaped by the same forces that are responsible for the codon pair preferences in genes of their hosts and posited that viral preferences reflect those of their host(s). The main goal of this study was, therefore, to determine the level of similarity in codon pairs used by human viruses and their host and to identify factors that are responsible for the selection of codon pairs in viruses. We analyzed the protein coding sequences of a large number of human viruses and discovered that the encoding of viruses is influenced only marginally by codon pair preferences of the host. We observed, however, that the encoding of viral genes mimics the dinucleotide bias of their hosts to a large extent. Furthermore, the similarity in codon pair preferences between human viruses and their host can be explained largely by CpG and, to a lesser extent, TpA suppression. Analysis of codon pair preferences in viruses from the *Reoviridae*, *Flaviviridae*, *Togaviridae*, and *Bunyaviridae* families showed that arboviruses do not use codon pairs that are overrepresented in the host arthropod species more often than viruses that are not transmitted by arthropods, confirming our previous observation that codon pair preferences of the

OPEN ACCESS
CellPress

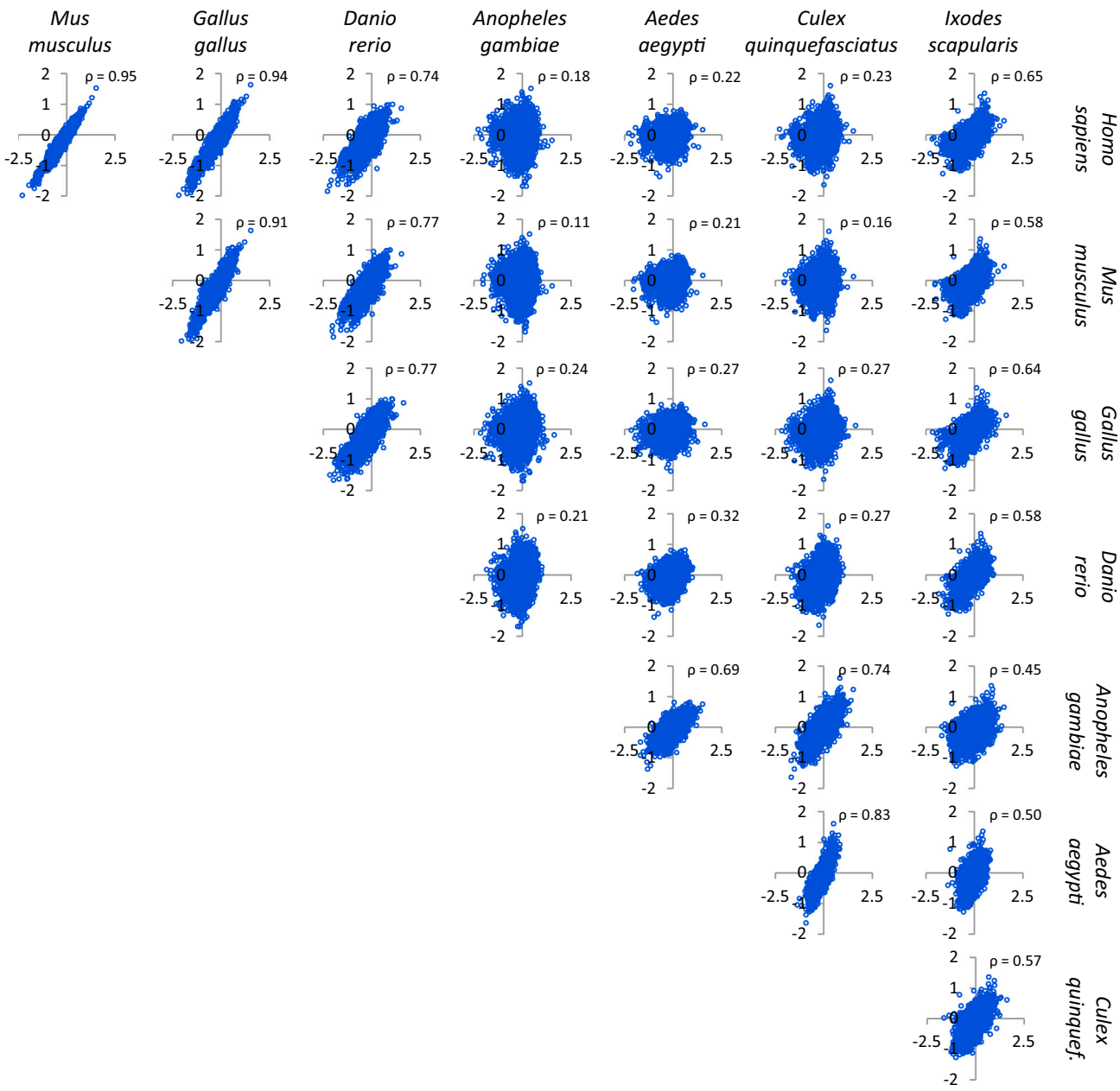


Figure 1. Correlation of CPSs among Selected Species

Species that are phylogenetically related have highly similar CPSs, suggesting that codon pair preferences reflect their common evolutionary history and selection forces that shape codon pairing in coding sequences.

host do not significantly influence the primary sequence and codon pair utilization of viral genes.

RESULTS

Codon Pair Preference in Different Species

First, we analyzed codon pair use in five vertebrates (human, pig, mouse, chicken, and zebrafish) and four arthropods (*Aedes aegypti*, *Anopheles gambiae*, *Culex quinquefasciatus*, and *Ixodes scapularis*). We used well annotated protein coding sequences to calculate the codon pair scores (CPSs) of all possible 3,721 codon pair combinations (Coleman et al., 2008). The CPS indicates whether a given codon pair is underrepresented relative to the expectation (<0) and, therefore, potentially avoided or overrepresented (>0) in the particular ORFeome.

A comparison of CPSs derived from different organisms (Figure 1; Figure S1) confirmed that closely related species have a similar codon pair bias (CPB) (Moura et al., 2007; Shen et al., 2015). For example, the CPSs of human, pig, and mouse are almost identical (Spearman's ρ 0.95), but even more evolutionary

distant species, such as the chicken and zebrafish, have CPSs significantly similar to those of the human (Spearman's ρ 0.94 and 0.74, respectively). Similarly, we detected high levels of correlation between CPSs derived from different mosquitoes (Figure 1; Figure S1) and noticed that codon pair preference is a stable property of a species because randomly selected subsets of the ORFeome, for example ORFeomes of two different chromosomes, produced almost identical CPSs.

Codon Pair Preferences in Human Viruses

We used human CPSs to determine whether viruses that infect humans have similar codon pair preferences as their host. We analyzed 92 (41 DNA and 51 RNA) viruses from all seven groups of viruses according to the Baltimore classification (Table S1). Most of the human viruses predominantly use codon pairs that have positive CPSs and are potentially preferred in the human, but the percentage of codon pairs with positive CPSs in viral genes is lower (50%–60%) than in human genes (65%) (Figure 2D). Moreover, when we quantified the level of codon pair under- or overrepresentation in ORFs, it became obvious that genes of human viruses contain codon pairs that have much lower CPSs than that of the human. As a result, ORFs of RNA and DNA viruses have a significantly lower CPB scores (mean of all CPSs in an ORF) than their human counterparts (Figures 2A–2C). Although the vast majority of human ORFs have CPB scores in the range of 0–0.2, the ORFs of RNA viruses have CPB scores ranging between −0.1 and 0.1, and ORFs of DNA viruses have even lower CPB scores (Figures 2A–2C).

When we calculated the overall CPB scores for viral ORFeomes, a similar picture emerged. We discovered that encoding in most of the human viruses is only marginally influenced by host codon pair preference (Figure 2E). Consequently, the CPB scores of many viruses were less than 0, and only in two cases (*Influenza C virus* and *BK polyomavirus)* did they match the CPB of the human ORFeome (0.075).

RNA Viruses

As stated above, human viruses from all five RNA classes were biased toward the use of codon pairs overrepresented in the human ORFeome, but the majority of viral ORFs had relatively low CPB scores in comparison with the host (Figure 2A). With the only clear exception represented by the viruses of the family *Togaviridae*, there are no obvious differences in CPB among different RNA virus families. Viruses belonging to the same family have varying CPB scores, but the range of CPB scores within each family—as in the entire group of RNA viruses—is narrow (<0.15 CPB). We did not observe any distinct differences in CPB between virus groups (e.g., single-stranded RNA [ssRNA] and double-stranded RNA [dsRNA] viruses), which suggests that there is no relationship between CPB and genome structure either.

DNA Viruses

In general, ORFs of DNA viruses have even lower CPB scores than those of RNA viruses (Figures 2A–2C and 2E). Similarly to RNA viruses, we did not find any evidence for a correlation between the structure of the genome and codon pair preferences in coding sequences. However, there was a clear negative correlation between genome size and overall CPB of DNA viruses (Figure 3). Viruses with the smallest genomes (~5 kb, polyomaviruses and parvoviruses) had clearly the highest CPB scores (0–0.08), papillomaviruses (~7.5 kb) and intermediate-sized adenoviruses (~35 kb) had lower CPB scores (−0.03–0.03), and herpesviruses and poxviruses (genomes >150 kb) had the lowest CPB scores (−0.07–0.03).

The Abundance of CpG Dinucleotides in Viral Genomes Correlates Negatively with CPB

To understand the possible factors that influence CPB in viruses, we examined whether the size or the nucleotide and dinucleotide composition of coding sequences correlate with the CPB of viruses (Figure 3). This analysis showed a clear negative correlation between genome size and CPB in DNA viruses: the larger the genome, the smaller the CPB (Figure 3). We did not detect any correlation between genome size and CPB in the case of RNA viruses. Intuitively, however, the G+C content (GC content) of RNA viruses negatively and clearly correlated with CPB ($R^2 = 0.46$). This was not the case for DNA viruses ($R^2 = 0.10$), suggesting that overall base composition influences CPB exclusively in RNA viruses.

The relative abundances (odds ratios) of dinucleotides deviate from the normal or expected distribution in a variety of genomes (Campbell et al., 1999; Karlin and Burge, 1995). Compared with other dinucleotides, TpA and CpG are the most underrepresented, whereas TpG, CpA, and CpT are the most overrepresented dinucleotides in the vertebrate genomes (Figure S2). Dinucleotide bias also appears to be linked with codon pair bias (Moura et al., 2007; Shen et al., 2015), but this relationship has not been thoroughly explored. For unknown reasons, CpG and TpA dinucleotides are also significantly suppressed in the genomes of most RNA and small DNA vertebrate viruses (Karlin et al., 1994).

The analysis of relative abundances of dinucleotides in coding sequences of human viruses revealed that CpG and TpA dinucleotides deviate the most from the mathematical prediction (Figure 2F), with CpG dinucleotides showing the highest level of suppression in most of the small DNA and all RNA viruses. TpA dinucleotides are underrepresented in most herpesviruses and small viruses. Unexpectedly, CpA and TpG dinucleotides are not only overrepresented in small DNA viruses, which might be explained by the cytosine methylation-deamination-mutation hypothesis (Bird, 1980), but also in the majority of RNA viruses. The relative abundance of CpG, TpG, and CpA dinucleotides generally conforms with random expectations in large DNA viruses (herpesviruses and poxviruses).

From the data presented in Figures 2D–2F, it is apparent that the relative abundance of CpG dinucleotides shows a strong negative correlation with the CPB in both DNA and RNA viruses (R^2 ~0.75; Figure 3). Therefore, the relative abundance of CpG dinucleotides plays a crucial role in determining the similarity in codon pair preferences between human viruses and their host.

Dinucleotide Bias Is the Main Force Responsible for Shaping Codon Pair Bias

To better understand the relationship of dinucleotide and codon pair bias, we analyzed the influence of five adjacent and ten non-adjacent nucleotide pairs in codon pairs on CPSs in vertebrates (Figure 4) and different mosquitoes (Figure S3). In the human and

Please cite this article in press as: Kunec and Osterrieder, Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias, Cell Reports (2016), http://dx.doi.org/10.1016/j.celrep.2015.12.011

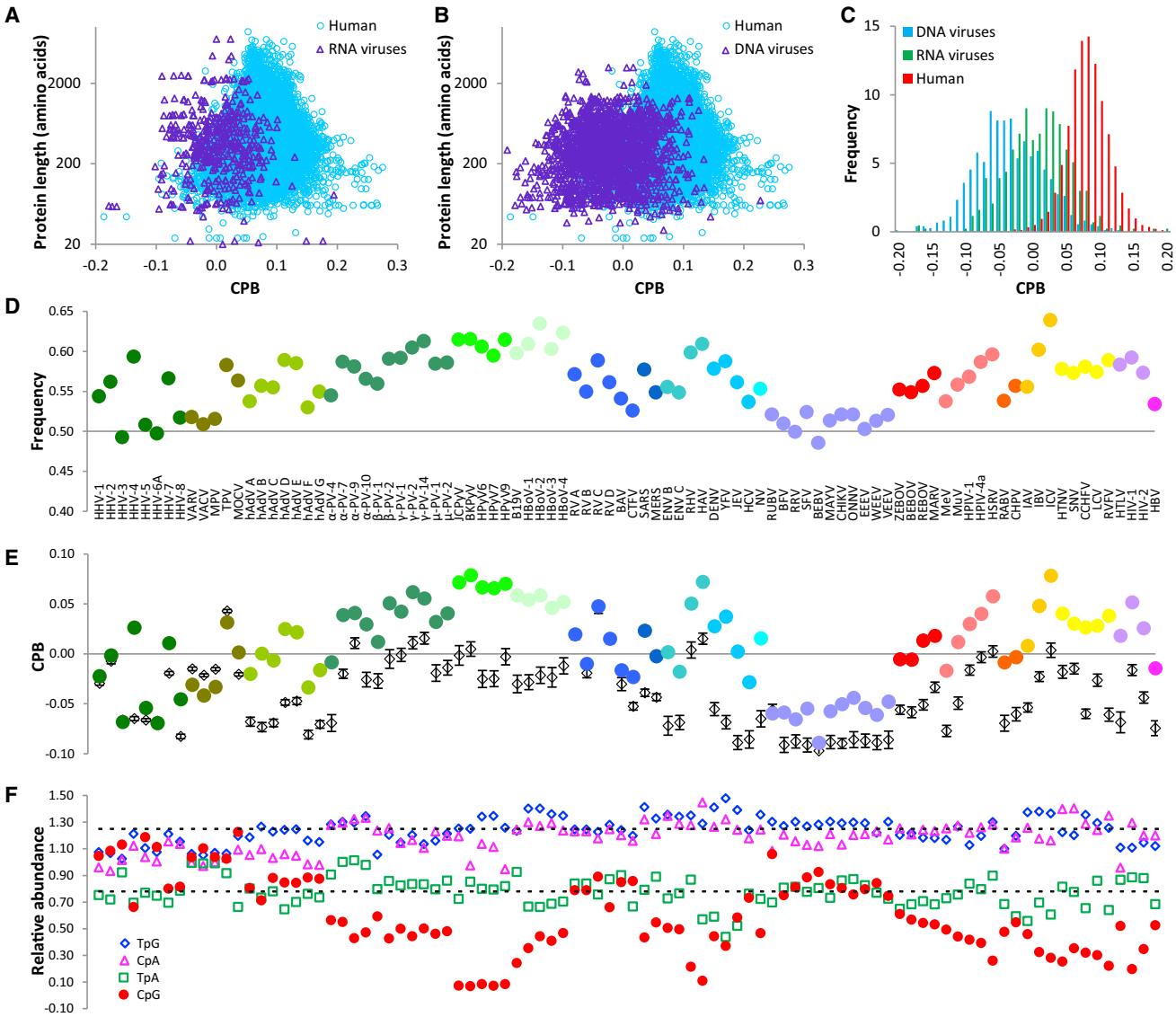


Figure 2. Codon Pair Bias and Dinucleotide Bias in Human Viruses

(A–C) The CPB scores of human and viral ORFs. The blue circles represent CPB scores of the 18,261 human ORFs. Purple triangles represent CPB scores of ORFs in RNA (A) and DNA (B) viruses. The CPB scores were calculated as a mean of CPSs of codon pairs present in the ORF. The CPB score of each ORF is plotted against its length. The majority of human ORFs have a positive CPB value (the CPB average of the entire human ORFeome = 0.075). In general, human ORFs have higher CPB scores than those of the viruses, and ORFs of RNA viruses have higher CPB scores than those of the DNA viruses. Also shown is the distribution of CPB scores in the human and human viruses (C).

(D) Frequency of codon pairs in protein coding sequences of viruses that are overrepresented in the human ORFeome (CPS > 0). Viruses are color-coded by family.

(E) The CPB scores of WT (dots) and randomized (diamonds) virus ORFeomes. Error bars represent mean ± SD.

(F) Relative abundance (odds ratios) of TpG, CpA, TpA, and CpG dinucleotides in protein coding sequences of analyzed viruses. From data experience the odds ratios that are located outside of the interval of 0.78–1.25 (dashed lines) are considered to be of low (high) relative abundance compared with a random association of nucleotides.

Definitions of virus name abbreviations are provided in Table S1.

other vertebrates, the largest deviation from the random distribution can be seen on the overrepresentation of CpG and TpA dinucleotides at the codon boundary (position P3-A1) in the underrepresented codon pairs (Figure 4). In contrast, TpG, CpA, and CpT are frequently seen in overrepresented codon pairs (Figure 4). Therefore, as expected, dinucleotides at position P3-A1 have a decisive role in directing the codon pair bias in analyzed organisms.

Because the CPS appeared to be influenced by the relative abundance of dinucleotides present at the codon pair boundary, we grouped codon pairs into 16 groups according to the dinucleotides that they contain at the codon pair boundary (i.e., one

OPEN ACCESS
CellPress

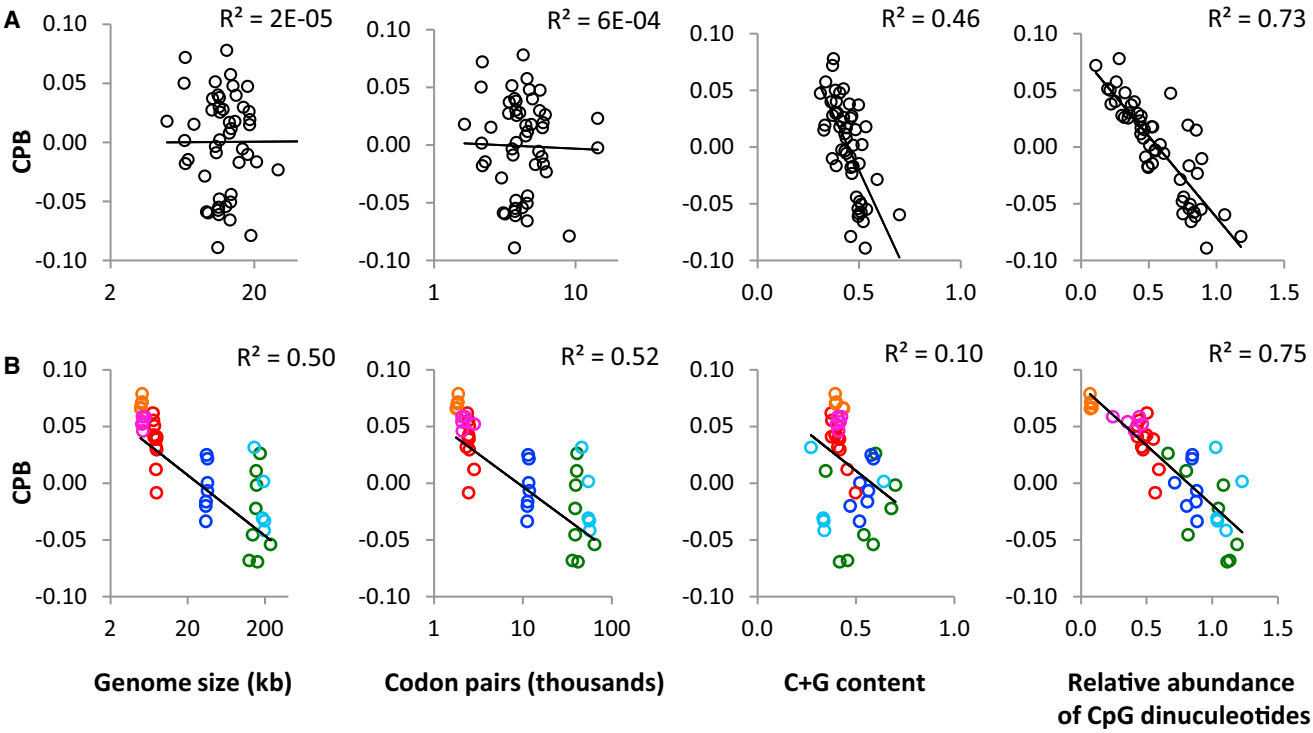


Figure 3. Analysis of Viral Genome Properties that Might Influence the Average Virus CPB Scores

(A and B) The average CPB scores of RNA (A) and DNA (B) viruses were correlated with the genome size, ORFeome size, C+G content, and relative abundance of CpG dinucleotides in the viral ORFeome. DNA viruses are color-coded by family: herpesviruses (green), poxviruses (light blue), adenoviruses (dark blue), papillomaviruses (red), polyomaviruses (orange), and parvoviruses (purple).

group would be the NNC-GNN codon pairs) and correlated the mean CPS of each group with the relative abundance of dinucleotides observed in coding sequences at the codon pair boundary (Figure 5). This experiment showed a significant correlation between the two factors.

Importantly, we also calculated CPSs for the remaining two non-coding (nonsense) reading frames utilizing the same strand in the human ORFeome. Again, we observed a high degree of correlation between CPS and dinucleotide bias in reading frame 3 (Figure 5), as we did between CPSs calculated for reading frames 1 and 3 (Spearman's ρ 0.60; Figure S4). This suggested that identical codon pairs have similar CPSs, although the CPSs had been calculated using highly dissimilar nucleotide sequences. Collectively, the above results clearly show that codon pair bias is a direct consequence of dinucleotide bias and that underrepresented codon pairs are not underrepresented because they are unfit for encoding proteins but are simply suppressed by forces that drive dinucleotide bias.

Random Reshuffling of Synonymous Codons in Viral ORFeomes

We eliminated the original codon pair ordering in the viral ORFs by random reshuffling of synonymous codons and then calculated CPB scores of the randomized ORFeomes. We expected that reshuffling would have a negligible effect on the CPB of viruses in which CpG and TpA dinucleotides are in the normal range. Conversely, we expected that viruses with considerable suppression of CpG dinucleotides will also have a high ΔCPB (difference between the CPB of the wild-type and reshuffled ORFeome) because reshuffling normalizes the relative abundance of dinucleotides at the codon pair boundary, and the level of normalization depends on the length and codon bias of the ORF.

According to our expectations, the overall CPB scores in viruses of the families *Poxviridae*, *Herpesviridae*, and *Reoviridae* that have normal abundances of CpG were not affected by randomization (Figure 2E; Table S1). CPB scores were reduced slightly in viruses that show moderate suppression of CpG dinucleotides. Unexpectedly, randomized ORFeomes of orthopoxviruses had an even higher CPB scores than the wild-type ORFeomes, indicating that random reshuffling created codon pairs that, on average, had higher CPSs than codon pairs in the wild-type ORFeome. This increased the level of similarity in codon pair preferences between the viruses and the hosts at the same time. In contrast, viruses with the highest suppression of CpG also had the highest ΔCPB between wild-type and randomized sequences. In other words, random reshuffling of synonymous codons in viruses that do not display suppression of CpG dinucleotides does not change the overall CPB score despite the complete permutation of coding sequences.

Arboviruses Do Not Mimic the Codon Pair Preferences of Their Arthropod Hosts

Arboviruses have the ability to replicate in both a vertebrate and arthropod host. The trade-off hypothesis predicts that

OPEN ACCESS
CellPress

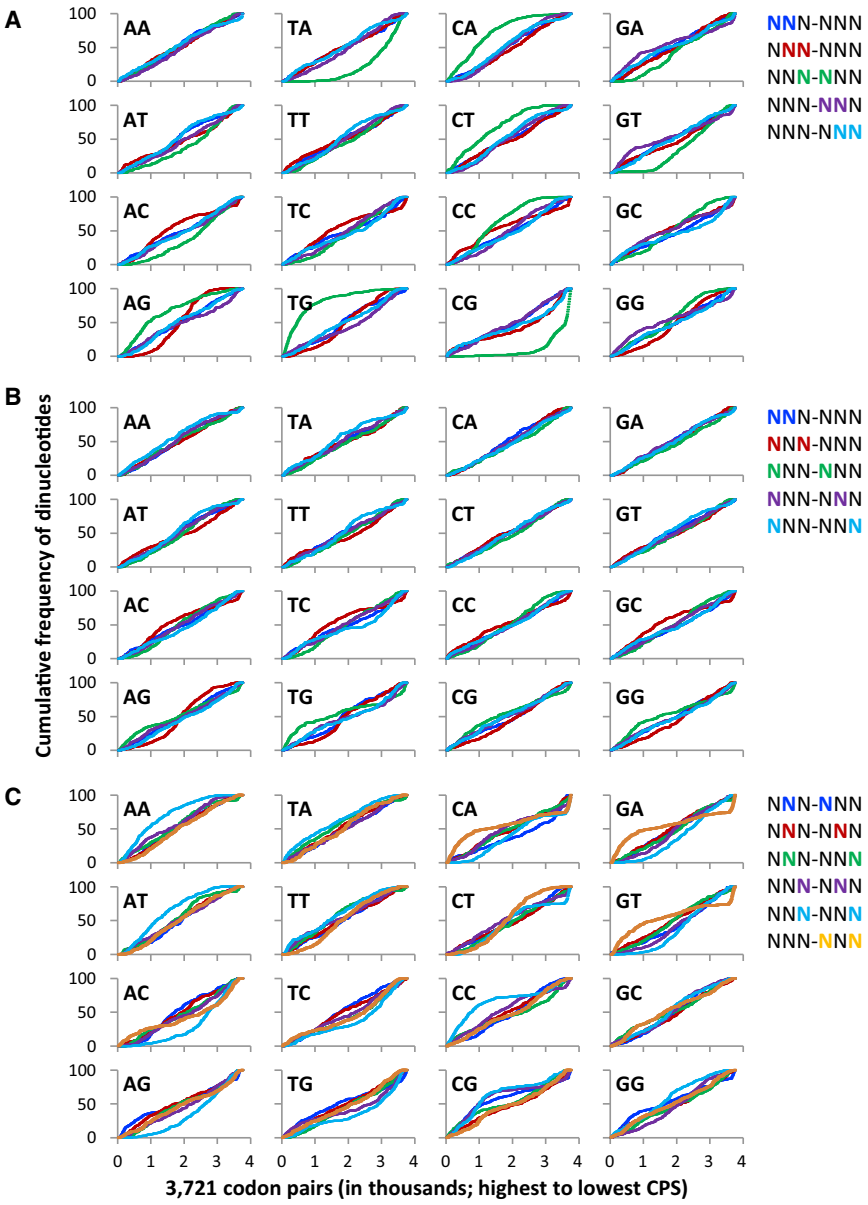


Figure 4. The Contribution of Adjacent and Nonadjacent Nucleotide Pair Combinations to Codon Pair Preferences in Protein Coding Sequences of *Homo sapiens*

(A–C) The contribution of adjacent (A) and nonadjacent (B and C) nucleotide pair combinations to codon pair preferences in protein coding sequences of the human. The 3,721 possible pairs of codons in protein coding sequences were sorted according to their CPSs, and the cumulative frequency of 15 possible nucleotide pair combinations was then calculated in the sorted array of codon pairs to identify nucleotide pairs that associate with the codon pair preferences. If codon pair preferences were not affected by the nature of nucleotide pairs in codon pairs, then the relationship between the cumulative frequency of nucleotide pairs and the rank number of codon pairs would have a linear trend. Codon pair preferences are affected mainly by the combination of certain dinucleotides that span the codon pair junction.

constraining evolution in one host species diminishes fitness in the other (Vasilakis et al., 2009). It has also been suggested that arboviruses use codon pairs that are overrepresented in both hosts to support efficient protein production in either environment (Shen et al., 2015). Our objective was to determine whether codon pairs in arboviruses are actively selected according to the codon pair preferences of their alternating hosts.

Arboviruses comprise a large and polyphyletic group of viruses that are transmitted between vertebrate hosts by hematophagous arthropod vectors (Hanley and Weaver, 2008). Almost all arboviruses are RNA viruses that belong to the *Reoviridae*, *Flaviviridae*, *Togaviridae*, and *Bunyaviridae* families. However, not all viruses from these families are arboviruses. On the contrary, all four RNA families contain species that do not require an arthropod for transmission, implicating that the arthropod-borne lifestyle evolved several times independently in different groups of viruses (Hanley and Weaver, 2008). We took advantage of this fortuitous situation and included in our analysis sister viruses from the non-arboviral genera because potential adaptations toward codon pair preferences of the vector should be more pronounced in arboviruses and discernible by comparison with viruses that are not arthropod-borne.

In total, we analyzed 159 different viruses from four different RNA families: *Reoviridae* (53 viruses), *Flaviviridae* (62 viruses), *Togaviridae* (30 viruses), and *Bunyaviridae* (14 viruses) (Table S2). From the family *Reoviridae*, we analyzed 27 arboviruses from the genera *Seadornavirus*, *Orbivirus*, and *Coltivirus*; 13 animal viruses from the genera *Rotavirus* and *Orthoreovirus*; nine arthropod-transmitted plant viruses from the genera *Phytoreovirus*, *Oryzavirus*, and *Fijivirus*; and four insect-specific viruses from the genus *Cypovirus*. From the *Flaviviridae*, we analyzed 55 viruses from the genus *Flavivirus* and eight mammalian viruses of the genera *Pestivirus*, *Hepacivirus*, and *Pegivirus*. From the *Togaviridae*, we analyzed human *Rubella virus*, the sole member of the genus *Rubivirus*, and 29 arboviruses from the genus *Alphavirus*. Finally, from the *Bunyaviridae*, we analyzed eight arboviruses from the genera *Orthobunyavirus*, *Nairovirus*, and *Phlebovirus* and three mammalian and three plant viruses from the genera *Hantavirus* and *Tospovirus*, respectively. To discover how well different viruses are codon pair optimized for different arthropod vectors, we calculated CPB scores of their ORFs using the CPSs derived from the ORFeomes of three different model mosquitoes (*Aedes aegypti*, *Anopheles gambiae*, and *Culex quinquefasciatus*) and a tick vector (*Ixodes scapularis*) (Table S3).

OPEN ACCESS
CellPress

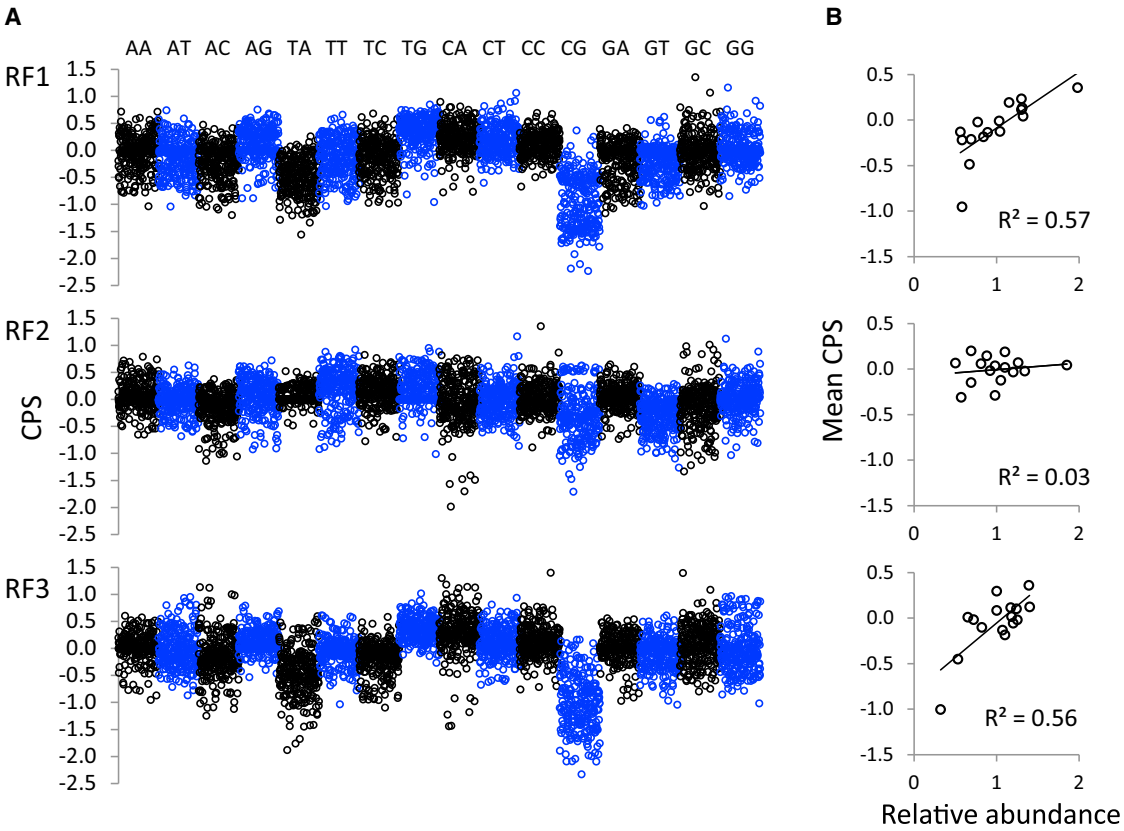


Figure 5. Codon Pair Scores Are Influenced by the Nature of the Dinucleotides that Occupy the Codon Pair Boundary

(A) Distribution of CPSs calculated for three possible reading frames in the human ORFeome. Each dot represents a CPS. Codon pairs were grouped into 16 groups (each containing 256 codon pairs) according to the dinucleotide from the codon pair boundary (indicated above each group of CPSs).

(B) For each group of codon pairs we calculated the mean CPS and correlated it with the relative abundance of dinucleotides corresponding to the dinucleotides present at the codon pair boundary.

Although CPB scores are distributed differentially in different virus families, we did not detect significant differences in CPB scores between arboviruses and non-vectored viruses within individual virus families. The CPB scores calculated using arthropod CPSs were not significantly higher in the arthropod-borne or arthropod-specific viruses than in the sister vertebrate viruses that do not replicate in arthropods (Figure 6). This finding can be well illustrated using the *Flaviviridae* family, which contains a large number of taxonomically recognized species that belong to one of four genera as outlined above (Moureau et al., 2015). Although the genus *Flavivirus* contains arboviruses, the remaining three genera contain exclusively non-vectored animal viruses. In addition, the genus *Flavivirus* also contains viruses that are hosted exclusively by insects or mammals. The arboviruses are further divided into mosquito-borne and tick-borne flaviviruses. The insect viruses are either classic insect-specific flaviviruses, which appear to have never acquired the ability to replicate in vertebrates, or insect-specific-like flaviviruses, which likely evolved from mosquito-borne viruses (Blitvich and Firth, 2015; Moureau et al., 2015). The viruses that are exclusively hosted by mammals (rodents and bats) are evolutionarily related to either tick- or mosquito-borne viruses (Moureau et al., 2015).

In general, we did not find evidence to suggest that codon pairs in the analyzed viruses are subject to selection for codon pairs that are preferred in their respective arthropod hosts. For example, the CPB scores of tick-borne flaviviruses, calculated using the CPSs of *Ixodes scapularis*, are not significantly higher than CPB scores of other viruses of the same family when using the tick CPSs (Figure 6). Similarly, the CPB scores of the mosquito-specific flaviviruses (insect-specific and insect-specific-like groups) calculated using different mosquito CPSs do not differ from the CPB scores of arboviruses that replicate in both hosts or viruses that are maintained in mammals without a vector (Figure 6). In addition, flaviviruses that infect arthropods did not contain more codon pairs that are overrepresented in their respective arthropod host than vertebrate-specific viruses.

The only association between the viral host and CPB can be observed within the insect-specific and the insect-specific-like groups of flaviviruses, which appear to have lower CPB scores relative to the human than most other flaviviruses (Figure 6). Mosquitoes, unlike vertebrates, do not display underrepresentation of CpG and overrepresentation of TpG and CpA dinucleotides in their genomes (Figure S2). The relative abundance of CpG dinucleotides is much lower in vertebrate-only than in insect-specific flaviviruses, suggesting that the two groups of

OPEN ACCESS
CellPress

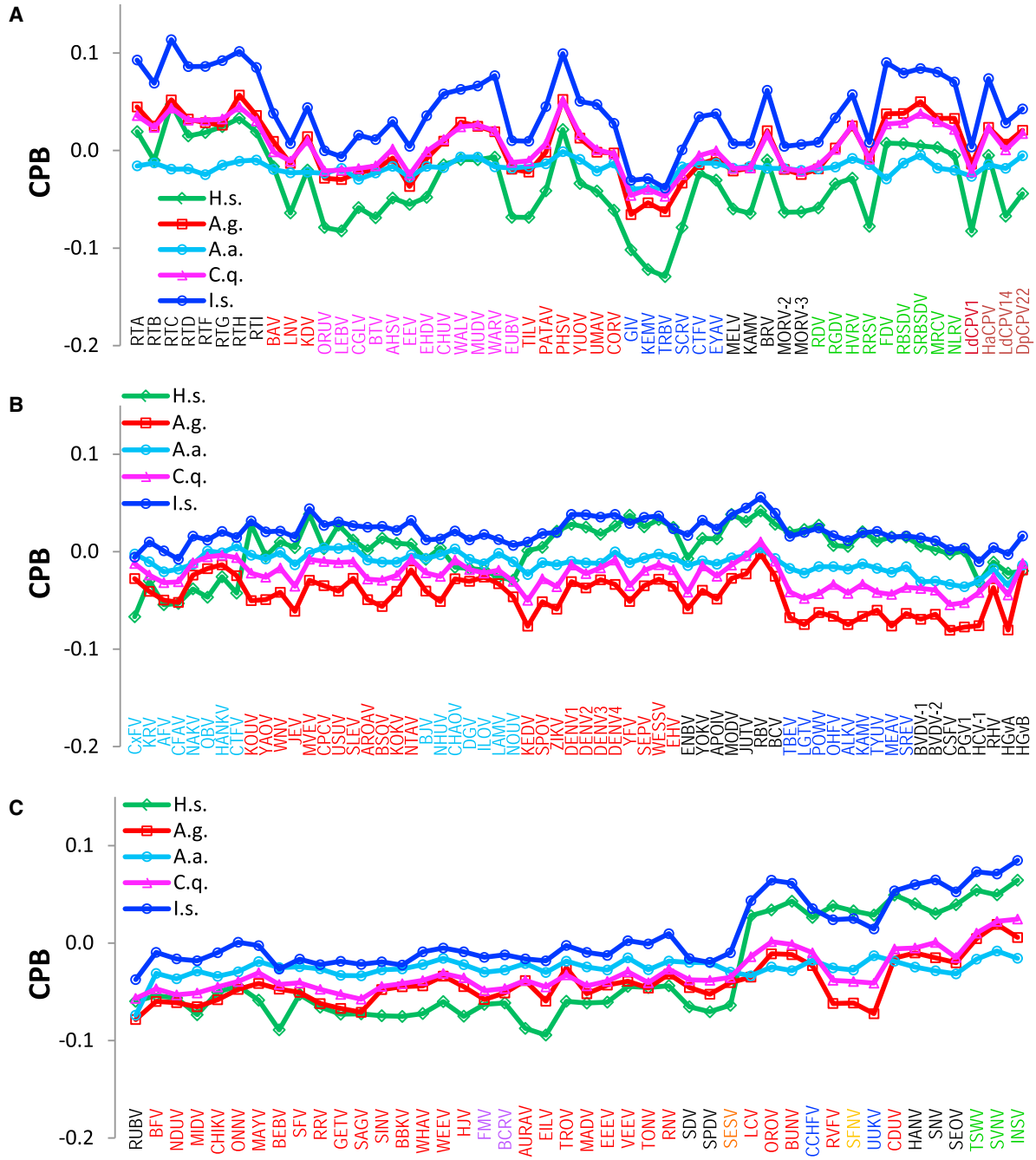


Figure 6. Analysis of Codon Pair Preferences in Vertebrate, Insect, Plant, and Arthropod-Borne Viruses from the families *Reoviridae*, *Flaviviridae*, *Togaviridae*, and *Bunyaviridae*

(A–C) Analysis of codon pair preferences in vertebrate, insect, plant, and arthropod-borne viruses from the families *Reoviridae* (A), *Flaviviridae* (B), and *Togaviridae* and *Bunyaviridae* (C).The average virus CPSs were calculated using the CPSs of selected species: *Homo sapiens* (H.s.), *Anopheles gambiae* (A.g.), *Aedes aegypti* (A.a.), *Culex quinquefasciatus* (C.q.), and *Ixodes scapularis* (I.s.). CPB > 0 means that coding sequences of a virus are mainly composed of codon pairs that are overrepresented in the particular species. Definitions of virus name abbreviations are provided in Table S2.

viruses are exposed to different selection pressures and that the pressure is imposed by the respective hosts (Lobo et al., 2009). Therefore, the slight underrepresentation of codon pairs that have high human CPSs in the insect-specific flaviviruses is not a result of selection for codon pairs that are preferred in the mosquitoes but, rather, a consequence of nonexistent selection against CpG dinucleotides in viral sequences. Accordingly, the use of CpG, on average, is higher in insect-specific flaviviruses,

and this causes the decrease of their human CPB scores. This interpretation is supported by the fact that insect-specific flaviviruses, on average, use more CpG dinucleotides than insect-specific-like viruses, and these more than the dual-host flaviviruses from which they evolved (Figure S5).

Similar conclusions can be derived by analysis of CPB scores in viruses from the other three virus families (Figure 6). We determined the relationship between the relative abundance of CpG and human-based CPB (Figure S6) and the relative abundances of 16 possible dinucleotides for all analyzed viruses (Figure S5).

To further corroborate our observation that the selection of codon pairs in arthropod-infecting viruses is not influenced by the codon pair preferences of their arthropod hosts, we analyzed the level of under/overrepresentation in codon pairs that are used by arthropod-infecting viruses from the *Reoviridae*, *Flaviviridae*, *Togaviridae*, and *Bunyaviridae*. The results of the analysis are exemplified in Figure 7. We discovered that none of the analyzed arboviruses show a bias toward codon pairs that are overrepresented in their respective mosquito vectors. In addition, analysis of the mosquito-specific flaviviruses *Culex flavivirus* (CxFV) and *Aedes flavivirus* (AFV) and other insect-specific flaviviruses (data not shown) showed that their ORFeomes are not biased toward codon pairs that have high CPSs in their respective mosquito host (Figure 7). We discovered that arthropod-borne flaviviruses and bunyaviruses show a bias toward codon pairs that are overrepresented in vertebrates, which was expected because viruses from these families show significant suppression of CpG dinucleotides in their genomes (Figure S5). In contrast, the togaviruses do not preferentially use codon pairs that have higher CPSs in their respective vertebrate hosts.

DISCUSSION

The goal of our study was to determine the level of similarity in codon pair preferences between human viruses and their host. We expected that the selection in viruses would largely reflect host codon pair preferences. The hypothesis was based on previous observations that increasing the number of codon pairs that are underrepresented in the host caused robust attenuation of several RNA viruses (Coleman et al., 2008; Le Nouën et al., 2014; Mueller et al., 2010; Shen et al., 2015; Wang et al., 2015; Yang et al., 2013).

Unexpectedly, we discovered that codon pair preferences in viruses that infect (primarily) humans only correlate very weakly with those of their hosts. On average, only about 50%–60% of codon pairs used by human viruses are overrepresented in the human ORFeome (Figure 2). In addition, codon pairs used by viruses have much lower CPSs than those that are used by the hosts. As a result, ORFeomes of many human viruses have negative CPB scores (Figure 2), which means that the level of underrepresented codon pairs in their genomes is greater than that of overrepresented codon pairs. The similarity in codon pair bias was higher in viruses that have small genomes. This, however, is caused by suppression of CpG dinucleotides in the viral genomes rather than by the actual selection of codon pairs that have high CPSs in the host and would, therefore, hypothetically be preferred for encoding proteins.

We also analyzed the protein coding sequences of a large number of arboviruses because their maintenance in nature requires replication in phylogenetically distant hosts (Erwin and Davidson, 2002) with very different codon pair preferences (Figure 1; Table S3). It has been proposed that attenuation by SAVE is a consequence of the presence of underrepresented codon pairs in protein coding sequences that do not support efficient protein production or processing (Coleman et al., 2008; Shen et al., 2015). Therefore, we surmised that, if overrepresented codon pairs were indeed preferred, then arboviruses should use codon pairs that are overrepresented in both hosts. In contrast to a previous study (Shen et al., 2015), we did not find evidence that would suggest that codon pairs in arboviruses are biased toward codon pairs preferred in both hosts. Although some of the analyzed arboviruses show a bias for codon pairs that are overrepresented in the vertebrate hosts, these viruses also display a marked suppression of CpG dinucleotides. On the other hand, protein coding sequences of viruses such as *Chikungunya virus* or *O'nyong'nyong virus*, which display only moderate suppression of CpG dinucleotides in their genomes (relative abundance 0.81 and 0.76, respectively), contain only very few codon pairs that are overrepresented in the vertebrate host (52%), which results in very low CPB scores (−0.05 and −0.04, respectively) (Figure 2). Therefore, as in the human viruses, the similarity in codon pair preferences of arboviruses can simply be explained by suppression of CpG dinucleotides, which concomitantly increases the number of codon pairs that are overrepresented in vertebrates.

Codon pair preferences were analyzed previously in arthropod-borne *Dengue virus 2* (DENV-2); *Rift Valley fever virus* (RVFV), a bunyavirus infects mosquitoes and sheep; and *Maize fine streak virus* (MFSV), a leafhopper-transmitted nucleorhabdovirus that replicates in an insect and a plant (Shen et al., 2015). Similar to our results, the data presented for DENV-2 and RVFV show that these viruses are biased for codon pairs that are overrepresented in the vertebrate host but not in the respective mosquito vector (Shen et al., 2015). The bias in analyzed arboviruses for overrepresented codon pairs of the corresponding vertebrates was identified only in a small set of codon pairs that are used with higher frequency. Again, the bias is observed only in arboviruses (such as DENV and RVFV) that display suppression of CpG dinucleotides.

Compared with other dinucleotides, the relative abundance of TpA and CpG deviates the most from the normal or expected distribution in a variety of organisms. Although TpA dinucleotides are underrepresented in the genomes of almost all species, CpG dinucleotides are underrepresented only in the nuclear DNA of plants and vertebrates as well as in the mitochondrial genomes of all metazoan species, be they invertebrates or vertebrates (Cardon et al., 1994). TpA is the least energetically stable dinucleotide (Breslauer et al., 1986) and TpA dinucleotides are often found in regions that require binding of proteins for bending and unwinding of the DNA double helix (e.g., TATA boxes or replication origins) (Karlin and Ladunga, 1994). Therefore, restriction of TpA use might reduce the improper binding of regulatory factors (Karlin and Burge, 1995; Karlin and Ladunga, 1994). In addition, TpA shows greater suppression in DNA that is destined for expression from mRNA in the cytosol (Beutler

OPEN ACCESS
CellPress

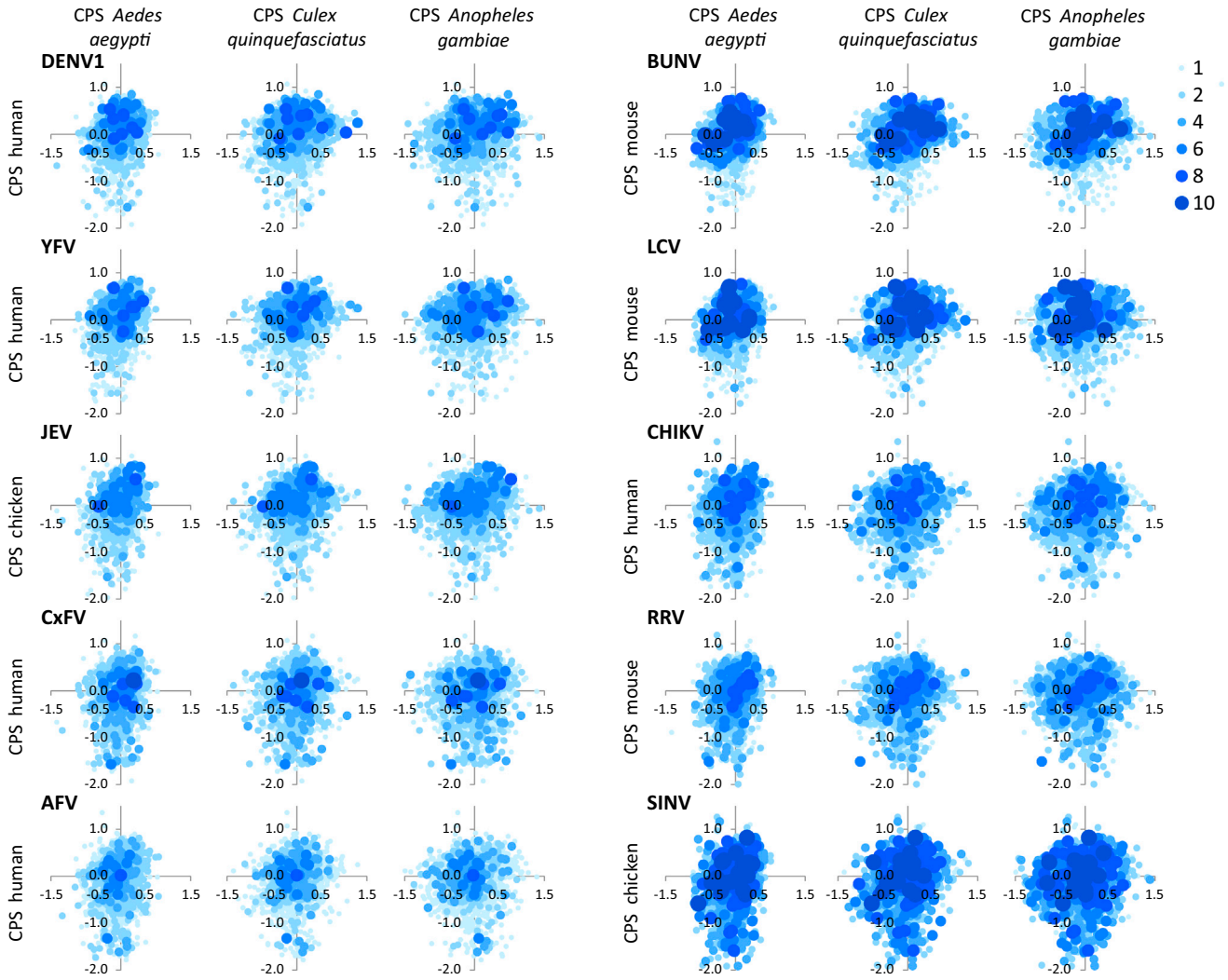


Figure 7. The Frequencies and CPSs of Codon Pairs Used by the Virus in the Arthropod (and Vertebrate) Host

The CPSs are shown as dots. The size and shade of a dot signifies the frequency of a codon pair in the virus genome. DENV-1, *Dengue virus 1* (*Aedes aegypti*, human); YFV, *Yellow fever virus* (*Aedes aegypti*, primates); JEV, *Japanese encephalitis virus* (*Culex* spp., birds); CxFV, *Culex flavivirus* (*Culex spp.*); AFV, *Aedes flavivirus* (*Aedes* spp.); BUNV, *Bunyamwera virus* (*Aedes aegypti*, rodents); LCV (*Aedes triseriatus*, rodents), CHIKV (*Aedes* spp., primates and rodents); RRV, *Ross river virus* (*Culex* and *Aedes* spp., mammals and birds); SINV, *Sindbis virus* (*Culex* spp., birds). The primary arthropod (and vertebrate) hosts are indicated in parentheses. DENV-1, YFV, JEV, CxFV, and AFV are flaviviruses; BUNV and LCV are bunyaviruses; and CHIKV, RRV, and SINV are togaviruses. Definitions of virus name abbreviations are provided in Table S2.

et al., 1989). This suppression may be accounted for by the fact that TpA dinucleotide is present in two canonical stop codons (TAA and TAG) and that UpA dinculeotides in RNA are preferential targets for ribonucleases (Beutler et al., 1989). Avoidance of TpA reduces the risk of nonsense mutations and increases the stability of RNA. CpG suppression is traditionally explained by the methylation-deamination-mutation hypothesis, where deaminated and unrepaired 5-methylcytosine mutates into thymine and results in conversion of CpG/CpG into TpG/CpA (Bird, 1980).

Previous studies have shown that SAVE not only increases the number of codon pairs that are underrepresented in coding sequences of the target host but that, when done using vertebrate CPSs, inadvertently increases the number of CpG and, to lesser degree, also TpA dinucleotides in recoded sequences because codon pairs that contain CpG and TpA dinucleotides at codon position P3-A1 are among the most underrepresented codon pairs (Table S3; Atkinson et al., 2014). We also assessed the alteration of 16 possible dinucleotide frequencies in recoded genes by analyzing codon pair deoptimized sequences that have been described in two recent studies on *Human respiratory syncytial virus* (HRSV) (Le Nouën et al., 2014) and DENV-2 (Shen et al., 2015). In both studies, three different viral regions were recoded (Min A, Min B, and Min L of HRSV and E, NS3, and NS5 of DENV-2). As expected, recoding changed the frequencies of TpA, CpG, TpG, and CpA dinucleotides (Figure S7). Although

the number of TpA dinucleotides in recoded genes increased only moderately (30%–60% increase over the wild-type), the number of CpG dinucleotides increased dramatically (140%–360% increase in the case of DENV-2 and 480%–630% in HRSV). The most prominent reduction of dinucleotide frequencies involved TpG and CpA dinucleotides (20%–60% decrease). This analysis seems to lend further support to our interpretation that attenuation of vertebrate viruses by SAVE is not caused by increasing non-preferred codon pairs but, rather, by increasing underrepresented dinucleotides.

The exact molecular mechanisms responsible for attenuation by SAVE or, rather, an increase of the CpG and TpA dinucleotides are still unknown. Two major, mutually not exclusive hypotheses propose explanations for the attenuation by SAVE and the CpG/TpA increase. The first theory posits that underrepresented codon pairs create unfavorable conditions for protein production, processing, or folding and that the decreased protein production is directly responsible for virus attenuation (Coleman et al., 2008). The alternative theory suggests that the cause of attenuation is to be found in the increased number of CpG (and TpA) dinucleotides, which are recognized by an as yet uncharacterized self-non-self recognition system that stimulates enhanced innate immune responses to such recoded viruses (Atkinson et al., 2014; Greenbaum et al., 2009).

Based on our findings, we put forward the hypothesis that codon pairs that contain CpG or TpA dinucleotides at the codon pair boundary might not be underrepresented in protein coding sequences because they are less fit for encoding proteins but, simply, because CpG and TpA dinucleotides show the highest level of suppression (relative abundance 0.48 and 0.56, respectively). A staggering 94% and 98% of codon pairs that contain TpA and CpG dinculeotides at position P3-A1 are underrepresented in the human ORFeome (Table S3). Similarly, the GpT dinucleotide has the third-lowest relative abundance (0.79; Figure S2), and 78% of codon pairs with such a dinucleotide at the codon pair boundary have negative CPSs. Conversely, 89% and 85% of codon pairs with TpG and CpA dinucleotides at the codon pair boundary have positive CPSs. Therefore, suppression of CpG (and TpA) dinucleotides in the genomes of vertebrates also causes codon pairs with CpG and TpA at the codon pair boundary to become underrepresented. The logical implication of this conclusion is that that attenuation by SAVE is not caused by impaired gene decoding or protein production but, rather, by a different mechanism.

It is becoming clear that many extant RNA vertebrate and plant viruses evolved from insect viruses (Li et al., 2015; Marklewitz et al., 2015). During adaptation viruses often change their genome structure according to the genome features of their new hosts (Greenbaum et al., 2009). The most interesting adaptation is the suppression of CpG (and TpA) dinucleotides in the genomes of viruses that infect vertebrates. This can be well exemplified with members of the family *Flaviviridae*. Suppression of CpG dinucleotides is higher in vertebrate-specific than in the arthropod-borne viruses, and is virtually nonexistent in classic insect-specific viruses (Lobo et al., 2009). Suppression of CpG dinucleotides in these and many other small viruses occurs at all three possible codon locations, not only at the codon pair boundary. This, however, would be expected if the theory were correct that links underrepresented codon pairs with suboptimal protein production because codon pairs that contain CpG dinucleotides at the codon pair boundary are the most underrepresented codon pairs in vertebrate genomes. The interpretation is supported by the fact that almost all analyzed arboviruses show stronger suppression of CpG dinucleotides at codon position P1-P2 and P2-P3 than at position P3-A1 (data not shown), suggesting that elimination of CpG dinucleotides from codon positions P1-P2 and P2-P3 is at least as important for the virus as elimination of these dinucleotides from position P3-A1; i.e., at the codon boundary that would have the largest effect on codon pair usage.

The exact basis for attenuation by SAVE still remains unanswered. Because codon pair preferences and dinucleotide frequencies are intimately related, dissecting the effects of the two on virus fitness is rather difficult. A study that analyzed a library of echovirus mutants - in which either CpG and TpA frequency or the overall CPB was kept constant while the other feature was altered - demonstrated that the increase of CpG and TpA frequencies impaired virus fitness but that alternation of CPB without changing the CpG and TpA frequency did not (Tulloch et al., 2014). Therefore, the results of this study, although questioned (Shen et al., 2015), are in line with our observations, which also suggest that the basis of attenuation by SAVE is the increase of CpG dinucleotides in coding sequences. An alternate confirmation of this hypothesis could be achieved by recoding viruses so that the recoded viruses would have a lower CPB, a lower frequency of highly underrepresented codon pairs, but a higher number of CpG dinucleotides at the first and/or the second codon position than the parental viruses. Therefore, if our conclusions are correct, then replication of such viruses and their overall fitness should be impaired in comparison with their parents. Conversely, if the alternate hypothesis is correct (and the presence of CpG dinucleotides in underrepresented codon pairs is not a prerequisite for attenuation), then it should be straightforward to put it to the test by altering the codon pair preferences of viruses for hosts that do not show suppression of CpG dinucleotides in their genomes. Because mosquitoes, and insects in general, do not display suppression of CpG dinucleotides, it should be possible to deoptimize codon pairs of insect-specific or arthropod-borne viruses for their insect hosts without altering the level of CpG (and TpA) dinucleotides in coding sequences. Such modified viruses should be highly attenuated in their insect hosts.

Both attenuation by SAVE and increase of CpG/TpA dinucleotides appear to be breakthrough technologies that might result in the production of very efficient and safe vaccines. The data presented here indicate that the basis for attenuation by SAVE is the increase of CpG dinucleotides in coding sequences of viruses. This means that viruses that were engineered by SAVE might not be weakened per se but that they are attenuated because they induce stronger immune responses (Tulloch et al., 2014). If this is true, then the genetic stability, safety, and efficacy of such attenuated virus mutants must be studied and tested exhaustively before this technology can be used for the development of animal or human vaccines.

The safety of codon pair deoptimized viruses should be studied in outbred populations, where one would expect differential

Please cite this article in press as: Kunec and Osterrieder, Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias, Cell Reports (2016), http://dx.doi.org/10.1016/j.celrep.2015.12.011

levels of attenuation depending on the genetic background of infected animals. It should also be possible to identify pathways used in the recognition of viral sequences with elevated TpA and CpG frequencies because attenuated viruses should still cause disease in individuals with compromised recognition mechanisms. Such experimental confirmation of our predictions is currently being performed but is beyond the scope of this work.

EXPERIMENTAL PROCEDURES

Retrieval of Protein Coding Sequences

The entire sets of protein coding sequences were retrieved from the NCBI Consensus CDS (CCDS) (https://www.ncbi.nlm.nih.gov/CCDS), NCBI Genome (http://www.ncbi.nlm.nih.gov/genome/), or the VectorBase databases (https://www.vectorbase.org/) using the Biomart tool (see Table S4 for details).

Calculation of Codon Pair Scores

To determine codon pair biases in coding sequences, we developed algorithms that calculate CPSs and CPB scores exactly as described by Coleman et al. (2008). The CPS is defined as the natural logarithm of the ratio of the observed over the expected number of occurrences of a particular codon pair in all protein coding sequences of a species. The expected number of codon pair occurrences estimates the number of codon pairs to be present if there is no association between the codons that form the codon pair. It is also calculated to be independent of codon bias and amino acid frequency (Coleman et al., 2008). A negative CPS value means that a particular codon pair is underrepresented, whereas a positive CPS value indicates that a particular codon pair is overrepresented in the analyzed protein coding sequences. Codon pairs that are equally under- or overrepresented have a CPS equidistant from 0. Mammals share essentially the same codon pair bias, which can be different among phylogenetically distant species (Mueller et al., 2010). We calculated CPS for each of the 3,721 possible codon pairs (61 × 61 codons) using only validated protein coding ORFs. We considered ORFs valid when they started with an ATG codon, ended with an in-frame stop codon, and had no internal stop codons or undetermined nucleotides. In the final set of ORFs, we included only the longest of the alternative splicing variants.

We used a core set of consistently annotated protein coding sequences (CDS) from the CCDS database to calculate species-specific CPSs for human (database name CPPDS15) and mouse (database name CCDS16). Similarly, we used the entire sets of protein coding sequences to calculate CPSs for pig, chicken, zebrafish, *Aedes aegypti*, *Anopheles gambiae*, *Culex quinquefasciatus* (*pipiens*), and *Ixodes scapularis*. All calculated CPSs are provided in Table S3.

Using the CPSs, we then calculated CPB scores for each analyzed ORF (ORFeome) as an average of the CPSs of all codon pairs present in each ORF (ORFeome). To determine whether codon pair ordering in the WT ORF was a result of chance, we randomly reshuffled synonymous codons in WT ORFs and generated a set of 30 randomized ORFs from each ORF. Random reshuffling removed codon pair preferences but preserved codon bias; i.e., all ORFs contained exactly the same codons. For each set of the reshuffled ORFs, we calculated the mean CPB (CPB random [rnd]), the SD, and the probability that codon ordering in the WT ORF was a result of random chance. For each ORF, we also calculated ΔCPB as a difference of CPB of the WT ORF (CPB WT) and the CPB rnd. Similarly, we calculated ΔCPB for each virus ORFeome. The CPB score of the wild-type ORF or ORFeome provides general information on the use of codon pairs that are over- or underrepresented relative to the human ORFeome. The comparison between the CPB scores of the WT and the reshuffled ORF or ORFeome provides information on the ordering of "available" codons in the particular virus.

Assessment of Dinucleotide Relative Abundances

We assessed the dinucleotide biases (relative abundances) in coding sequences using the odds ratio measure $\rho XY = fXY/fXfY$, where fXY denotes the observed frequency of the dinucleotide XY and fXfY the product of the frequency of the nucleotides X and Y in a sequence (Burge et al., 1992). As a conservative criterion, we considered dinucleotides XY with $\rho XY < 0.78$ (> 1.25) of low (high) relative abundance because each ρXY occurs with the probability of less than 0.001 for sufficiently long (~20 kb) random sequences.

Identification of Dinucleotide Pairs that Influence Codon Pair Scores

We sorted all 3,721 possible codon pairs in a descending order by their CPSs (from highest to lowest) and calculated the distribution of different nucleotide pairs in CPS-sorted codon pairs. Including the stop codons, there are 4,096 (64 × 64) possible codon pairs, and there are 256 different codon pairs that contain a particular type of a dinucleotide at a particular position (e.g., NNC-GNN). We analyzed the contribution of 16 possible dinucleotides in five possible adjacent nucleotide pair types (P1-P2, P2-P3, P3-A1, A1-A2, and A2-A3) and ten possible non-adjacent nucleotide pair types (P1-P3, P1-A1, P1-A2, P1-A3, P2-A1, P2-A2, P2-A3, P3-A2, P3-A3, and A1-A3) on codon pairing preferences in different species. To visualize the distribution of a particular nucleotide pair in a sorted array of codon pairs, we plotted the cumulative frequency of dinucleotides against the rank number of a particular codon pair.

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.12.011.

AUTHOR CONTRIBUTIONS

Conceptualization, D.K. and N.O.; Software, D.K.; Writing – D.K. and N.O.; Visualization, D.K.; Funding Acquisition, D.K. and N.O.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (OS 143/5-1 to N.O.).

Received: July 21, 2015
Revised: November 3, 2015
Accepted: November 23, 2015
Published: December 24, 2015

REFERENCES

Atkinson, N.J., Witteveldt, J., Evans, D.J., and Simmonds, P. (2014). The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. Nucleic Acids Res. *42*, 4527–4545.

Beutler, E., Gelbart, T., Han, J.H., Koziol, J.A., and Beutler, B. (1989). Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc. Natl. Acad. Sci. USA *86*, 192–196.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. *8*, 1499–1504.

Blitvich, B.J., and Firth, A.E. (2015). Insect-specific flaviviruses: a systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization. Viruses *7*, 1927–1959.

Breslauer, K.J., Frank, R., Blöcker, H., and Marky, L.A. (1986). Predicting DNA duplex stability from the base sequence. Proc. Natl. Acad. Sci. USA *83*, 3746–3750.

Burge, C., Campbell, A.M., and Karlin, S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. USA *89*, 1358–1362.

Campbell, A., Mrázek, J., and Karlin, S. (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc. Natl. Acad. Sci. USA *96*, 9184–9189.

Cardon, L.R., Burge, C., Clayton, D.A., and Karlin, S. (1994). Pervasive CpG suppression in animal mitochondrial genomes. Proc. Natl. Acad. Sci. USA *91*, 3799–3803.

Coleman, J.R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., and Mueller, S. (2008). Virus attenuation by genome-scale changes in codon pair bias. Science *320*, 1784–1787.

Erwin, D.H., and Davidson, E.H. (2002). The last common bilaterian ancestor. Development *129*, 3021–3032.

Greenbaum, B.D., Rabadan, R., and Levine, A.J. (2009). Patterns of oligonucleotide sequences in viral and host cell RNA identify mediators of the host innate immune system. PLoS ONE *4*, e5969.

Gutman, G.A., and Hatfield, G.W. (1989). Nonrandom utilization of codon pairs in Escherichia coli. Proc. Natl. Acad. Sci. USA *86*, 3699–3703.

Hanley, K.A., and Weaver, S.C. (2008). Arbovirus evolution. In Origin and Evolution of Viruses, *Chapter 16*, Second Edition, E.D.R.P.J. Holland, ed. (London: Academic Press), pp. 351–391.

Karlin, S., and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. *11*, 283–290.

Karlin, S., and Ladunga, I. (1994). Comparisons of eukaryotic genomic sequences. Proc. Natl. Acad. Sci. USA *91*, 12832–12836.

Karlin, S., Doerfler, W., and Cardon, L.R. (1994). Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J. Virol. *68*, 2889–2897.

Le Nouën, C., Brock, L.G., Luongo, C., McCarty, T., Yang, L., Mehedi, M., Wimmer, E., Mueller, S., Collins, P.L., Buchholz, U.J., and DiNapoli, J.M. (2014). Attenuation of human respiratory syncytial virus by genome-scale codon-pair deoptimization. Proc. Natl. Acad. Sci. USA *111*, 13169–13174.

Li, C.X., Shi, M., Tian, J.H., Lin, X.D., Kang, Y.J., Chen, L.J., Qin, X.C., Xu, J., Holmes, E.C., and Zhang, Y.Z. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. eLife *4*, 4.

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

Marklewitz, M., Zirkel, F., Kurth, A., Drosten, C., and Junglen, S. (2015). Evolutionary and phenotypic analysis of live virus isolates suggests arthropod origin of a pathogenic RNA virus family. Proc. Natl. Acad. Sci. USA *112*, 7536–7541.

Moura, G., Pinheiro, M., Silva, R., Miranda, I., Afreixo, V., Dias, G., Freitas, A., Oliveira, J.L., and Santos, M.A. (2005). Comparative context analysis of codon pairs on an ORFeome scale. Genome Biol. *6*, R28.

Moura, G., Pinheiro, M., Arrais, J., Gomes, A.C., Carreto, L., Freitas, A., Oliveira, J.L., and Santos, M.A. (2007). Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. PLoS ONE *2*, e847.

Moureau, G., Cook, S., Lemey, P., Nougairede, A., Forrester, N.L., Khasnatinov, M., Charrel, R.N., Firth, A.E., Gould, E.A., and de Lamballerie, X. (2015). New insights into flavivirus evolution, taxonomy and biogeographic history, extended by analysis of canonical and alternative coding sequences. PLoS ONE *10*, e0117849.

Mueller, S., Papamichail, D., Coleman, J.R., Skiena, S., and 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*, 9687–9696.

Mueller, S., Coleman, J.R., Papamichail, D., Ward, C.B., Nimnual, A., Futcher, B., Skiena, S., and Wimmer, E. (2010). Live attenuated influenza virus vaccines by computer-aided rational design. Nat. Biotechnol. *28*, 723–726.

Shen, S.H., Stauft, C.B., Gorbatsevych, O., Song, Y., Ward, C.B., Yurovsky, A., Mueller, S., Futcher, B., and Wimmer, E. (2015). Large-scale recoding of an arbovirus genome to rebalance its insect versus mammalian preference. Proc. Natl. Acad. Sci. U.S.A *112*, 4749–4754.

Tulloch, F., Atkinson, N.J., Evans, D.J., Ryan, M.D., and Simmonds, P. (2014). RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. eLife *3*, e04531.

Vasilakis, N., Deardorff, E.R., Kenney, J.L., Rossi, S.L., Hanley, K.A., and Weaver, S.C. (2009). Mosquitoes put the brake on arbovirus evolution: experimental evolution reveals slower mutation accumulation in mosquito than vertebrate cells. PLoS Pathog. *5*, e1000467.

Wang, B., Yang, C., Tekes, G., Mueller, S., Paul, A., Whelan, S.P., and Wimmer, E. (2015). Recoding of the vesicular stomatitis virus L gene by computer-aided design provides a live, attenuated vaccine candidate. MBio *6*, 6.

Yang, C., Skiena, S., Futcher, B., Mueller, S., and Wimmer, E. (2013). Deliberate reduction of hemagglutinin and neuraminidase expression of influenza virus leads to an ultraprotective live vaccine in mice. Proc. Natl. Acad. Sci. USA *110*, 9481–9486.