Available online at www.sciencedirect.com

SCIENCE DIRECT®

Virus Research 92 (2003) 1–7

Virus Research

www.elsevier.com/locate/virusres

The extent of codon usage bias in human RNA viruses and its evolutionary origin

Gareth M. Jenkins, Edward C. Holmes *

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Received 5 June 2002; received in revised form 16 October 2002; accepted 18 October 2002

Abstract

Revealing the determinants of codon usage bias is central to the understanding of factors governing viral evolution. Herein, we report the results of a survey of codon usage bias in a wide range of genetically and ecologically diverse human RNA viruses. This analysis showed that the overall extent of codon usage bias in RNA viruses is low and that there is little variation in bias between genes. Furthermore, the strong correlation between base and dinucleotide composition and codon usage bias suggested that mutation pressure rather than natural (translational) selection is the most important determinant of the codon bias observed. However, we also detected correlations between codon usage bias and some characteristics of viral genome structure and ecology, with increased bias in segmented and aerosol-transmitted viruses and decreased bias in vector-borne viruses. This suggests that translational selection may also have some influence in shaping codon usage bias.

Keywords: Codon usage bias; Mutation pressure; Base composition; Dinucleotide; Translational selection

1. Introduction

Despite the importance of codon usage bias as an indicator of the forces shaping genome evolution, little is known about the extent and origin of this bias in RNA viruses. This sits in contrast to organisms such as bacteria, yeast, *Drosophila* and mammals, where codon usage bias has been studied in much greater detail, revealing extensive variation both within and among genomes (reviewed in Mooers and Holmes, 2000).

In general, codon usage bias may be the product of mutation pressure and/or natural selection for accurate and efficient translation. For example, in *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila* and *Arabidopsis thaliana*, highly expressed genes have a strong selective preference for codons with a high concentration of the corresponding acceptor tRNA molecule, whereas genes expressed at lower levels display a more uniform pattern of codon usage (Grantham et al., 1981; Gouy and Gautier, 1982; Sharp et al., 1986; Powell and Moriyama, 1997; Akashi and Eyre-Walker, 1998). In contrast, mutation pressure has been shown to be the dominant factor shaping both codon usage bias and base composition in mammalian genomes (Wolfe et al., 1989; Sharp et al., 1993; Francino and Ochman, 1999). In these species the extent of codon bias depends on the chromosomal location of each gene which is characterised by a particular mutation bias; for example, genes located in G+C rich regions of chromosomes preferentially utilize G+C ending codons. The most likely explanation for the lack of translational selection in mammals is that their relatively small population sizes mean that genetic drift will dominate the evolutionary dynamics of mutations that only differ marginally in fitness. Given that mutation rates in RNA viruses are very much higher than those in life-forms with DNA genomes (Drake and Holland, 1999), it is important to determine whether mutation pressure is also the main determinant of codon usage bias in these organisms. Finally, codon usage may also be influenced by an underlying bias in dinucleotide usage. As a case in point, CpG doublets in vertebrates occur at one fifth of the expected frequency due to methylation mutation

* Corresponding author. Tel.: +44-1865-271282; fax: +44-1865-310447

E-mail address: edward.holmes@zoo.ox.ac.uk (E.C. Holmes).

0168-1702/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0168-1702(02)00309-X

(Bird, 1986), and these doublets are now known to play a major role in determining chromatin structure and gene activation (Kundu and Rao, 1999).

Thus far, studies of codon usage bias in viruses have tended to focus on a few specific cases, rather than investigating the general causes of bias across a wide range of virus families. For example, the human immunodeficiency virus (HIV) has a marked codon usage bias due to its strong preference for the A nucleotide, which may reach frequencies of up to 0.40 (Hemert and Berkhout, 1995). Striking biases in base composition have also been described in pneumoviruses, where the overall G+C content is less than 0.33 in all cases (Pringle and Easton, 1997), and rubella virus, which has a genomic G+C content of 0.70 (Karlin et al., 1994). Nearly all RNA viruses are deficient in the dinucleotide CpG, although the reasons why are uncertain (Karlin et al., 1994). Studies on DNA viruses, which exhibit much lower rates of nucleotide substitution, have revealed a relationship between codon usage and tRNA availability in papillomavirus (Zhou et al., 1999), and codon usage may play a key role in regulating latent versus productive infection in Epstein-Barr virus (Karlin et al., 1990). In contrast, in nucleopolyhedroviruses codon usage appears to be simply a consequence of uneven base composition (Levin and Whittome, 2000).

Clearly, a better knowledge of codon usage bias in RNA viruses is essential to understanding the processes governing their evolution, particularly the overall role played by mutation pressure. Further, such information is relevant to understanding the regulation of viral gene expression and also to vaccine design where the efficient expression of viral proteins may be required to generate immunity (Hass et al., 1996). Herein, we describe the patterns of codon usage bias in a wide range of RNA viruses. Because codon usage bias varies extensively within and among host species, and it is unclear how this might affect codon choice in RNA viruses, we chose to focus on human RNA viruses alone in order to minimise the effects of host factors on codon bias. We also investigate the key evolutionary determinants of codon usage bias, particularly the respective contributions made by mutation pressure and natural selection.

2. Materials and methods

2.1. Sequences

Codon usage bias was measured in the 50 RNA viruses listed in Table 1. Three data sets were used in each case: the complete genomic coding region (excluding non-coding regions) and the regions encoding the RNA polymerase and nucleocapsid protein taken separately, since all viruses contain these genes. The average number of codons per sequence in each data set was 3297, 1172 and 427, respectively. Regions containing overlapping reading frames were excluded from all data sets. The viruses analysed are as follows (strain name, where available, followed by GenBank accession number are given in parentheses): Coxsackievirus A9 (Griggs; D00627), Enterovirus 71 (TW/2272/98; AF119795), Hepatitis A virus (HAF-203; AF268396), Poliovirus type 3 (23127; X04468), Rhinovirus type 89 (HRV89; M16248), Hepatitis E virus (Hyderabad; AF076239), Norwalk virus (BS5; AF093797), Astrovirus type 1 (Oxford; L23513), Dengue-1 virus (PDK-13; AF180818), Dengue-2 virus (S1; NC_001474), Dengue-3 virus (H87; M93130), Dengue-4 virus (NC_002640), GB virus C (G05BD; AB003292), Hepatitis C virus (HCV-A; AJ000009), Japanese encephalitis virus (GP78; AF075723), Murray Valley virus (MVE-1-51; AF161266), West Nile virus (2741; AF206518), Western tick-borne encephalitis virus (Neudoerfl; U27495), Yellow fever virus (Trinidad 79A; AF094612), Eastern equine encephalitis virus (U01034), O'nyong-nyong virus (SG650; AF079456), Ross River virus (NB5092; M20162), Rubella virus (Cendehill; AF188704), Sindbis virus (Edsbyn 82-5; M69205), Venezuelan equine encephalitis virus (83U434; U55362), Western equine encephalitis virus (71V-1658; AF214040), Mokola virus (1835157), Rabies virus (Nishigahara; AB044824), Vesicular stomatitis virus (Indiana; J02428), Measles virus (Edmonston; K01711), Mumps virus (Jeryl Lynn; AF201473), Parainfluenza-3 virus (D84095), Respiratory syncytial virus (S2 ts1C; U39661), Ebola virus (Mayinga; AF086833), Marburg virus (Popp; 450908), Influenza A virus (Akita/1/94, Beijing/32/92; U71132, U71128, U71136, U26830, U71144, U71140, U65564, U65670 for segment 1-8 respectively), Influenza B virus (Yamagata/16/88, Victoria/3/85; AF102006, AF101989, AF102023, X13553, AF100396, L49385, X67013, AF100378 for segment 1-8 respectively), Influenza C virus (AA-pi, JJ/50, California/78; U20228, M28060, M28062, K01689, M17700, M22038 for segment 1-7 respectively), Bunyamwera virus (X14383, M11852, D00353 for segment 1-3 respectively), Hantaan virus (H8205, Q32; D25531, AB030232, AB027097 for segment 1-3 respectively), La Crosse virus (74-32813; U12396, D10370, K00610 for segment 1-3 respectively), Rift Valley fever virus (ZH548, MM12; X56464, M11157, X53771 for segment 1-3 respectively), Sin Nombre virus (NM R11; L37902, L37903, L37904 for segment 1-3 respectively), Lassa fever virus (GA391; U73034, X52400 for segment 51, 2, respectively), Lymphocytic choriomeningitis virus (WE; AF004519, M20869 for segment 1–2 respectively), Rotavirus (KU, K8, 1076, IGV-S, TW941022; AB022765, AB022766, AB022767, D90260, AB022770, AB022769, D00325, AF190171, AB022772, 6009574, AF044355), HIV type-1 (VI850; AF077336), HIV type-2 (BEN;

Table 1
Observed and expected codon usage bias and base composition for 50 human RNA viruses

Virus	L_{AA}[a]	$G+C_{3S}$[b]	N_C[c]	N_C[d]	N_C[e]	N_C[e]	N_C[f]
Picornaviridae							
Coxsackievirus A9	2201	0.49	55.6	60.1	55.5	58.0	58.2
Enterovirus 71	2193	0.48	56.6	60.4	58.0	58.6	58.0
Hepatitis A virus	2227	0.26	38.9	43.4	37.4	39.5	41.0
Poliovirus type 3	2206	0.47	54.2	58.7	50.4	56.9	56.8
Rhinovirus type 89	2164	0.25	45.9	44.1*	43.2	44.7	45.5*
Caliciviridae							
Hepatitis E virus	2902	0.60	48.2	49.3*	49.9	45.9	48.9*
Norwalk virus	2522	0.44	56.4	57*	55.0	54.5	54.5*
Astroviridae							
Astrovirus type 1	2226	0.39	54.8	55.4*	52.4	56.4	53.6*
Flaviviridae							
Dengue-1 virus[‡]	3391	0.42	50.0	54.7	47.8	50.7	53.0
Dengue-2 virus[‡]	3388	0.42	48.6	54.2	48.9	38.3	52.0
Dengue-3 virus[‡]	3390	0.43	49.4	55.1	48.6	56.0	52.5
Dengue-4 virus[‡]	3387	0.44	50.9	56.9	50.8	55.8	53.5
GB virus C	2836	0.65	53.7	56.0	53.3	48.5	53.2*
Hepatitis C virus	3014	0.68	51.9	54.4	56.0	56.5	51.2*
Japanese encephalitis virus[‡]	3412	0.52	55.8	59.8	54.3	61.0	56.9
Murray Valley virus[‡]	3412	0.46	53.6	59.4	52.5	54.7	56.1
West Nile virus[‡]	3412	0.53	53.8	59.9	53.1	53.0	55.4
Western tick-borne encephalitis virus[‡]	3412	0.57	54.5	60.0	54.6	58.9	55.3*
Yellow fever virus[‡]	3411	0.50	53.1	60.8	51.1	48.5	54.5
Togaviridae							
Eastern equine encephalitis virus[‡]	3735	0.50	58.3	60.7	58.8	58.7	59.3
O'nyong-nyong virus[‡]	2512	0.48	54.3	55.9	52.6	48.9	56.2
Ross River virus[‡]	2480	0.57	56.3	59.4	57.0	60.0	57.7
Rubella virus[§]	2128	0.80	39.0	38.6*	36.9	37.1	39.2*
Sindbis virus[‡]	3760	0.57	55.3	58.4	57.8	56.0	56.7
Venezuelan equine encephalitis virus[‡]	2491	0.53	55.6	60.2	56.5	57.4	57.7
Western equine encephalitis virus[‡]	2466	0.51	56.8	60.0	59.2	58.1	58.1
Rhabdoviridae							
Mokola virus	3604	0.46	52.7	59.7	52.5	53.1	55.7
Rabies virus	3600	0.48	53.6	59.9	53.5	55.3	54.1
Vesicular stomatitis virus[‡]	3536	0.39	52.4	56.2	52.2	48.2	53.0*
Paramyxoviridae							
Measles virus[§]	4213	0.47	55.1	59.71	54.6	54.2	54.3
Mumps virus[§]	4305	0.36	54.3	53.8*	53.8	54.0	56.8
Parainfluenza-3 virus[§]	4211	0.26	43.8	42.6*	41.4	49.5	53.8*
Respiratory syncytial virus[§]	4361	0.27	44.3	47.4	41.0	46.9	45.0
Filoviridae							
Ebola virus	4761	0.38	55.3	55.1*	53.0	57.3	45.7
Marburg virus	4873	0.34	53.4	53.8*	50.3	54.2	55.5*
Orthomyxoviridae							
Influenza A virus[†§?]	4291	0.40	52.4	56.2	49.8	54.5	53.4*
Influenza B virus[†§]	4397	0.33	46.7	49.5	45.6	50.1	53.6
Influenza C virus[†§]	3919	0.27	43.0	45.9	42.0	44.7	49.34
Bunyaviridae							
Bunyamwera virus[†‡]	3802	0.29	45.2	45.7*	43.3	54.0	45.8
Hantaan virus[†§]	3715	0.32	48.8	48.4*	45.8	50.7	49.2
La Crosse virus[†‡]	3848	0.33	47.7	50.5	47.4	55.6	47.2
Rift Valley fever virus[†‡]	3865	0.44	51.8	57.2	51.4	61.0	45.9
Sin Nombre virus[†§]	3722	0.30	47.5	47.7*	45.5	52.7	47.0*
Arenaviridae							
Lassa fever virus [†§]	3377	0.41	50.0	55.2	48.6	50.59	50.8*
Lymphocytic choriomeningitis virus[†§]	3355	0.45	50.9	58.5	50.5	50.2	51.5
Reoviridae							

Table 1 (*Continued*)

Virus	L_{AA}[a]	G+C_{3S}[b]	N_C[c]	N_C[d]	N_C[e]	N_C[e]	N_C[f]
Rotavirus[†]	5735	0.21	40.3	40.0*	40.7	42.0	52.5
Retroviridae							
HIV type-1	2425	0.32	44.0	46.9	41.9	44.2	46.7
HIV type-2	2294	0.38	47.2	48.6*	42.6	49.4	48.3*
Human T-lymphotropic virus type-1	1997	0.52	52.4	51.5*	52.6	51.5	44.1
Human T-lymphotropic virus type-2	1946	0.55	50.2	48.6*	49.3	50.0	50.9*

[†], [‡], [§], refer to segmented, vector-borne and aerosol-transmitted viruses, respectively.

[a] Number of codons analysed in the complete coding region.

[b] G+C content at synonymous third codon positions.

[c] Codon usage bias for the complete coding region.

[d] Expected codon usage bias for the complete coding region correcting for uneven base composition.

[e] Codon usage bias for the regions encoding the RNA polymerase and capsid protein, respectively.

[f] Expected codon usage bias correcting for uneven base composition and dinucleotide bias.

* Indicates viruses where the actual codon usage bias was not significantly greater than the expected value.

NC_001722), Human T-lymphotropic virus type-1 (AF033817), Human T-lymphotropic virus type-2 (Y13051).

2.2. Codon usage analysis

Codon usage bias was measured using the effective codon usage statistic, N_C (Wright, 1990). The reported value of N_C is always between 20 (when only one codon is used for each amino acid) and 61 (when all codons are used equally). Although N_C values are not normally distributed, the sample of RNA viruses examined here is large so that t tests (one-tailed, $P < 0.05$ taken as significant), in addition to nonparametric tests, were used to compare the (arithmetic) mean codon usage between viruses differing in particular genomic or ecological characteristics. To determine the extent to which base composition (mutation bias) accounts for the patterns of codon usage bias observed, the N_C statistic for each virus sequence was compared to that obtained in 1000 simulated sequences of the same length and base composition (for the four nucleotides individually) at synonymous third codon positions (i.e. G+C_{3S}) in the original sequence. This approach was used because G+C_{3S} is a good indicator of the extent of base composition bias, which is less pronounced if all codon positions are considered, and the N_C statistic is independent of the base composition at nonsynonymous positions since it takes into account uneven amino acid usage.

3. Results

The extent of codon usage bias was determined in the complete genomic coding region of 50 genetically and ecologically diverse human RNA viruses (Table 1). The effective number of codons (N_C) used by these viruses was on average 50.9 and ranged from 38.9 (Hepatitis A virus) to 58.3 (Eastern equine encephalitis virus). Thirty three viruses had N_C values in the range 50–61, 15 in the range 40–50 and 2 between 38 and 40. Therefore, codon usage bias is in most cases slight, although some variation is evident.

Since viruses could in theory contain genes with contrasting biases in codon usage, for example due to local base composition differences, we also measured codon bias separately in the gene regions encoding the RNA polymerase and nucleocapsid proteins. This analysis revealed a similar distribution of codon usage bias measures to that observed previously (Table 1). For the region encoding the RNA polymerase, the average codon usage index was 50.0 and ranged from 36.9 (Rubella virus) to 59.2 (Western equine encephalitis virus). For the nucleocapsid region, the average N_C was 52.1 and ranged from 37.1 (Rubella virus) to 61.0 (Japanese encephalitis virus). Furthermore, there was a good correlation between the N_C values for the polymerase and nucleocapsid regions, $r = 0.75$ ($P < 0.000001$), indicating that codon usage bias does not deviate greatly between individual virus genes. However, average codon usage bias was higher in the RNA polymerase gene than the nucleocapsid gene for viruses that encode these proteins in separate open reading frames (e.g. Rabies virus) (mean difference = 2.66, paired t test and Wilcoxon rank test, $P < 0.05$), whilst there was no significant difference among viruses that encode these proteins in the same open reading frame.

Perhaps the most important question relating to the evolution of codon usage bias is whether it is controlled by mutation pressure or by natural selection. To choose between these evolutionary forces we compared actual codon usage bias with the bias expected under the null hypothesis that mutation pressure is the sole determinant. First, we compared G+C content at first and second codon positions (G+C_{12}) with that at synonymous third codon positions (G+C_{3S}) (Fig. 1). A highly significant correlation was observed ($r = 0.89$, $P < 0.000001$), indicating that patterns of base composition

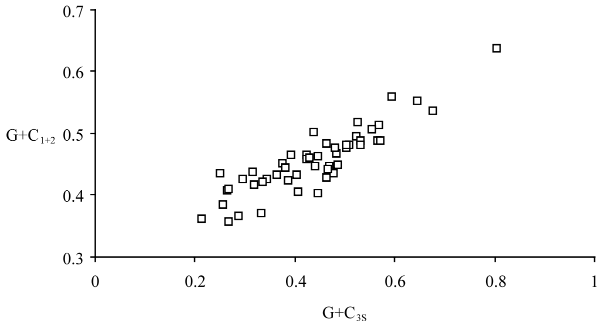

Fig. 1. Correlation between G+C content at first and second codon positions ($G+C_{12}$) with that at synonymous third codon positions ($G+C_{3S}$).

are most likely the result of mutation pressure, and not natural selection, since the effects are present at all codon positions. Next, for each virus, we plotted actual codon bias against both $G+C_{3S}$ and the expected N_C value if codon usage bias is solely due to biased base composition (i.e. G+C content). This revealed that the actual codon usage indices are close to the values expected from their G+C composition, although all are slightly lower (Fig. 2). As this deviation could be due to some uneven usage of G versus C and/or A versus U, we also compared the actual codon bias values for each virus with the N_C values calculated from randomised sequences of the same $G+C_{3S}$ base composition and overall length (Table 1). For 34 of the 50 viruses, the extent of codon bias observed was greater than that found in the randomised sequences ($P < 0.05$), indicating that uneven base composition alone cannot account for all of the codon usage bias in these viruses. Given that the average actual and expected codon usage indices are 50.9 and 53.8, respectively, and that a value of 61.0 represents no codon usage bias, we estimate that uneven base composition, and hence mutation pressure, accounts for on average 71.3% of the codon usage bias in these viruses (= 61 − average N_C expected from the randomised data sets /61 − average N_C observed in the actual data sets). Similar values were obtained when the RNA polymerase (77.5%) and nucleocapsid (85.3%) genes were considered separately. It is also possible that dinucleotide biases, which are independent of the overall base composition but still the result of differential mutation pressure, may shape patterns of codon usage, especially since these biases are present in virtually all viruses (Karlin et al., 1994). We therefore compared actual codon usage bias with that expected from randomised sequences of the same base and dinucleotide composition. In this case, uneven base and dinucleotide composition together accounted for an average of 88.1% of the observed codon usage bias in the complete coding region (Table 1).

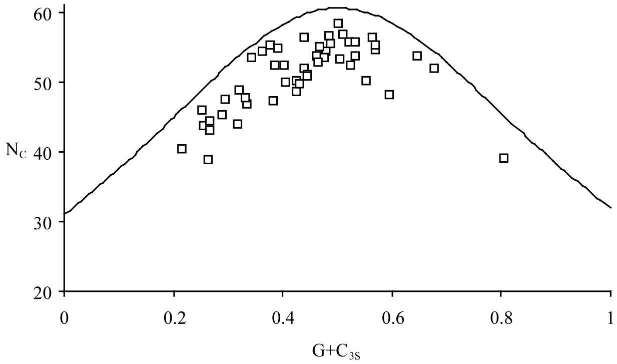

Fig. 2. Distribution of the codon usage index, N_C, and G+C content at synonymous third codon positions ($G+C_{3S}$). The curve indicates the expected codon usage if G+C compositional constraints alone account for codon usage bias.

As differences in mutation pressure explain most, but not all, the patterns of codon usage bias observed, we also sought to determine whether any viral genomic or ecological factors were correlated with the observed variation in codon bias. We therefore grouped codon usage indices for the complete coding region according to viral genome polarity, segmentation, presence of an envelope, genome length, transmission mode, duration of infection and type of disease caused. A number of noteworthy associations were found; average codon usage bias was higher in segmented ($N_C = 47.7$) than in non-segmented viruses (51.8) (*t* test, $P = 0.013$), higher in aerosol-transmitted viruses (47.8) than in other viruses (52.0) (*t* test, $P = 0.017$), but lower in vector-borne (52.8) than in non-vector-borne viruses (49.7) (*t* test, $P = 0.020$). All of these differences were also significant according to a nonparametric Wilcoxon rank test. However, in all cases the differences in bias are not large, and since this analysis did not correct for phylogenetic non-independence, it is possible that the results are confounded to some extent by shared evolutionary history. Finally, since a relationship between codon bias and gene length has been reported for some organisms, for example, a negative relationship in *Caenorhabditis*, *Drosophila* and *Arabidopsis* (Duret and Mouchiroud, 1999) and a positive relationship in *E. coli* (Moriyama and Powell, 1998), we also examined whether codon usage bias was correlated with the length of the polymerase and nucleocapsid genes. No significant relationship was found using either linear regression or Spearman's correlation ($P > 0.05$).

4. Discussion

This study of codon usage bias in human RNA viruses includes genetically and ecologically very diverse viruses representing most RNA virus families. Our analysis revealed that codon usage bias is slight in most cases, both for the complete coding region and the regions encoding two individual virus proteins. As a case in point, even hepatitis A virus, which had the greatest codon usage bias, still effectively uses two codons to encode each amino acid. The average N_C value of 50.9 among RNA viruses can be compared to those seen in other organisms such as *E. coli*, *S. cerevisiase* and *Drosophila melanogaster* where mean values of 45.0, 48.3 and 46.2, respectively, have been reported, and which all contain genes with high levels of codon usage bias (Powell and Moriyama, 1997). In the case of *Homo sapiens*, the average N_C value also probably lies close to 45 since the distribution of values for individual genes ranges uniformly from just under 30 to 61 (Wright, 1990). Thus, the general effects of codon usage bias in these organisms are more pronounced than those observed in human RNA viruses.

The general association between codon usage bias and base composition suggests that mutational pressure, rather than natural (translational) selection, is the most important general cause of patterns of codon usage in human RNA viruses, accounting for between ~71 and 85% of the bias observed, and almost 90% if the effects of dinucleotide bias are also incorporated. The importance of mutation over translational selection is further supported by the fact that RNA polymerase genes were often found to have greater codon usage bias than nucleocapsid genes, even though structural genes are generally expressed at much higher levels than non-structural genes.

Despite the evident importance of mutation pressure in determining patterns of codon usage bias in RNA viruses, statistically significant correlations were observed between codon bias and various characteristics of viral genome structure and ecology, most notably an increased bias in segmented and aerosol-transmitted viruses and a decreased bias in vector-borne viruses. Such correlations suggest that some aspects of codon choice in RNA viruses may be under the control of (relatively weak) translational selection, as it seems unlikely that unrelated viruses with a common property, such as aerosol transmission, would be subject to equivalent mutation pressures. In this respect, the finding that segmented RNA viruses had a higher codon usage bias than non-segmented RNA viruses is particularly notable, since one of the effects of reassortment is to increase codon bias by facilitating the action of natural selection (Kliman and Hey, 1994). It is also interesting that vector-borne RNA viruses had a lower codon usage bias than other RNA viruses. One possible explanation is that a low bias is advantageous to viruses that need to replicate efficiently in both insect and vertebrate cells, two very different cell types with potentially distinct codon preferences. However, as arboviruses have lower rates of nucleotide substitution than other RNA viruses (Jenkins et al., 2002), these viruses might have been expected to display *greater* codon usage bias, since an inverse relationship between rates of nucleotide substitution and codon bias has been documented in other organisms (Sharp and Li, 1987, 1989). Furthermore, experimental studies have suggested that insect and mammalian cells may constitute similar adaptive environments for arboviruses (Novella et al., 1999). Therefore, the reasons for the low codon bias in vector-borne RNA viruses are uncertain. Indeed, the lack of comprehensive codon usage tables for common vector species and of information about how the expression of viral genes relates to tRNA availability in either human or arthropod host cells makes it impossible to fully determine the importance of translational selection in RNA viruses.

Given the large population sizes of RNA viruses (typically 10^8–10^{11} within an infected host), it is perhaps surprising that the effects of selection are not more pronounced than documented here. Consequently, it may be that the mutation rates of RNA viruses are too high and the fitness effects of alternative codon choices too slight for translational selection to operate efficiently, or that effective population sizes, N_e, which refer to the number of viruses contributing progeny to the next generation, may often be sufficiently low for genetic drift to control molecular evolution (Leigh Brown, 1997). Alternatively, patterns of evolution at synonymous sites, and hence codon usage bias, may be influenced by RNA secondary structure, the effects of which are largely undetermined (Simmonds and Smith, 1999). Clearly, the underlying reasons for the correlation between codon usage bias and various aspects of RNA virus biology, and particularly whether they have a selective basis, need to be explored further.

Acknowledgements

This work was supported by the Wellcome Trust and the Royal Society.

References

Akashi, H., Eyre-Walker, A., 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Devel. 8, 688–693.

Bird, A.P., 1986. CpG-rich islands and function of DNA methylation. Nature 321, 209–213.

Drake, J.W., Holland, J.J., 1999. Mutation rates among RNA viruses. Proc. Natl. Acad. Sci. USA 96, 13910–13913.

Duret, L., Mouchiroud, D., 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, *Arabidopsis*. Proc. Natl. Acad. Sci. USA 96, 4482–4487.

Francino, H.P., Ochman, H., 1999. Isochores result from mutation not selection. Nature 400, 30–31.

Gouy, M., Gautier, C., 1982. Codon usage in bacteria: correlation with expressivity. Nucleic Acids Res. 12, 539–549.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., Mercier, R., 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9, 43–74.

Hass, J., Park, E., Seed, B., 1996. Codon usage limitation in the expression of HIV-1 envelope glycoprotein. Curr. Biol. 6, 315–324.

Hemert, F.J., Berkhout, B., 1995. The tendency of lentiviral open reading frames to become A-rich, constraints imposed by viral genome organisation and cellular tRNA availability. J. Mol. Evol. 41, 132–140.

Jenkins, G.M., Rambaut, A., Pybus, O.G., Holmes, E.C., 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J. Mol. Evol. 54, 152–161.

Karlin, S., Blaisdell, B.E., Schachtel, G.A., 1990. Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus-data and hypotheses. J. Virol. 64, 4264–4273.

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

Kliman, R.M., Hey, J., 1994. The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. Genetics 137, 1049–1056.

Kundu, T.K., Rao, M.R.S., 1999. CpG islands in chromatin organisation and gene expression. J. Biochem. 125, 217–222.

Leigh Brown, A.J., 1997. Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc. Natl. Acad. Sci. USA 94, 1862–1865.

Levin, D.B., Whittome, B., 2000. Codon usage in nucleopolyhedroviruses. J. Gen. Virol. 81, 2313–2325.

Mooers, A.Ø., Holmes, E.C., 2000. The evolution of base composition and phylogenetic inference. Trends Ecol. Evol. 15, 365–369.

Moriyama, E.N., Powell, J.R., 1998. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res. 26, 3188–3193.

Novella, I.S., Hershey, C.L., Escarmis, C., Domingo, E., Holland, J.J., 1999. Lack of evolutionary stasis during alternating replication of an arbovirus in insect and mammalian cells. J. Mol. Biol. 287, 459–465.

Powell, J.R., Moriyama, E.N., 1997. Evolution of codon usage bias in *Drosophila*. Proc. Natl. Acad. Sci. USA 94, 7784–7790.

Pringle, C.R., Easton, A.J., 1997. Monopartite negative strand RNA genomes. Semin. Virol. 8, 49–57.

Sharp, P.M., Li, W.H., 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4, 223–230.

Sharp, P.M., Li, W.H., 1989. On the rate of DNA-sequence evolution in *Drosophila*. J. Mol. Evol. 28, 398–402.

Sharp, P.M., Stenico, M., Peden, J.F., Lloyd, A.T., 1993. Codon usage: mutational bias, translational selection, or both? Biochem. Soc. Trans. 21, 835–841.

Sharp, P.M., Touhy, T.M.F., Mosurski, K.R., 1986. Codon usage in yeast, cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14, 5125–5143.

Simmonds, P., Smith, D.B., 1999. Structural constraints on RNA virus evolution. J. Virol. 73, 5787–5794.

Wolfe, K., Sharp, P.M., Li, W.H., 1989. Mutation rates differ among regions of the mammalian genome. Nature 337, 283–285.

Wright, F., 1990. The effective number of codons used in a gene. Gene 87, 23–29.

Zhou, J., Liu, W.J., Peng, S.W., Sun, X.Y., Frazer, I., 1999. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. J. Virol. 73, 4972–4982.