J Mol Evol (2006) 62:551–563
DOI: 10.1007/s00239-005-0221-1

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary Basis of Codon Usage and Nucleotide Composition Bias in Vertebrate DNA Viruses

Laura A. Shackelton,[1] Colin R. Parrish,[2] Edward C. Holmes[3]

[1] Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK
[2] J.A. Baker Institute, Department of Microbiology and Immunology, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA
[3] Center for Infectious Disease Dynamics, Department of Biology, Mueller Laboratory, The Pennsylvania State University, University Park, PA 16802, USA

Received: 17 September 2005 / Accepted: 20 December 2005 [*Reviewing Editor:* Dr. Nicolas Galtier]

Abstract. Understanding the extent and causes of biases in codon usage and nucleotide composition is essential to the study of viral evolution, particularly the interplay between viruses and host cells or immune responses. To understand the common features and differences among viruses we analyzed the genomic characteristics of a representative collection of all sequenced vertebrate-infecting DNA viruses. This revealed that patterns of codon usage bias are strongly correlated with overall genomic GC content, suggesting that genome-wide mutational pressure, rather than natural selection for specific coding triplets, is the main determinant of codon usage. Further, we observed a striking difference in CpG content between DNA viruses with large and small genomes. While the majority of large genome viruses show the expected frequency of CpG, most small genome viruses had CpG contents far below expected values. The exceptions to this generalization, the large gammaherpesviruses and iridoviruses and the small dependoviruses, have sufficiently different life-cycle characteristics that they may help reveal some of the factors shaping the evolution of CpG usage in viruses.

Key words: DNA viruses — Codon bias — Base composition — Mutation pressure — Natural selection — Dinucleotide bias — CpG

Correspondence to: Edward C. Holmes; *email:* ech15@psu.edu

Introduction

Even before the genetic code was deciphered, it was proposed that gene sequence evolution is not only influenced by fitness effects at the protein level, but also by the intrinsic nucleotide composition of the genome (Sueoka 1961). Once the redundancy of the genetic code was revealed, it became apparent that different organisms had evolved, along with "classical phenotypes," unique genomic signatures, or "genomic phenotypes" (Bernardi and Bernardi 1986). Of particular importance was the proposal that each species was subject to specific genomic pressures on base composition, in turn resulting in a distinctive bias in codon choice (Grantham et al. 1980), and that explaining these unique coding strategies "is the heart of the problem of molecular evolution" (Grantham et al. 1986).

More recent studies have revealed that patterns of codon usage bias and nucleotide composition within many cellular genomes are far more complex than previously imagined, and the factors shaping their evolution are still not entirely understood. In principle, biases in nucleotide composition and codon usage can result from natural selection and/or differential mutational pressure. In many organisms, such as *Escherichia coli*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Caenorhabditis elegans*, there is evidence that codons that use abundant tRNAs are selectively favoured, especially in highly expressed genes (Sharp et al. 1986; Powell and

Moriyama 1997; Gouy and Gautier 1982; Stenico et al. 1994). Codon selection of this type most likely functions to optimize translational speed and/or translational accuracy, although additional factors, such as transcription efficiency, mRNA secondary structure, and protein structure, can also exert selection pressures (Xia 1996; Zama 1990; Oresic and Shalloway 1998). In contrast, codon usage bias in mammals appears to be more strongly influenced by differential mutation pressure (Sharp et al. 1993), although some evidence for selection has also been observed (Smith and Eyre-Walker 2001; Chamary and Hurst 2004; Duan and Antezana 2003). Under the mutation pressure model, intrinsic differences in the propensity of genome replication to make specific mutational errors, usually depicted as the frequency of GC $\leftrightarrow$ AT changes, shapes overall patterns of base composition. Further, because the effective population sizes (N_e) of mammalian species are typically small, as are the selection coefficients (s) of most mutations at synonymous sites or in noncoding DNA (i.e., $N_e s << 1$), natural selection is usually unable to control the substitution dynamics of these mutations and they are fixed, or more usually lost, by genetic drift. Evidence for this "mutation-drift" theory in vertebrate genomes is that the nucleotide and codon bias of genes largely reflects their genomic location; for example, genes in GC-rich regions (so-called GC "isochores") are biased toward GC-ending codons (Sharp et al. 1993).

There have been few in-depth studies of codon and nucleotide usage biases among families of DNA viruses, although the biases of some eukaryotic viruses can differ substantially from those of their hosts (Strauss et al. 1996). For example, in the mammalian papillomaviruses it has been proposed that a codon usage bias different from the average seen in host genomes strongly influences both viral replication and gene expression (Zhao et al. 2003). Moreover, in the rapidly evolving human RNA viruses, one of the few groups of viruses for which codon bias data have been compiled, mutation pressure seems to be the main force shaping codon usage, accounting for 71–85% of the observed bias (Jenkins and Holmes 2003). Indeed, in one in-depth analysis of the nidovirales, neither translational selection nor gene length was found to have an effect on codon usage (Gu et al. 2004). Although RNA viruses possess large effective population sizes, it is possible that their mutation rates are so high that they prevent natural selection from working efficiently on codon choice (Jenkins and Holmes 2003).

Codon usage can also be strongly influenced by underlying biases in dinucleotide frequency, which differs greatly among organisms. Specifically, after accounting for dinucleotide biases, the proportion of codon usage bias explained by mutation pressure often increases, as seen in human RNA viruses (Jenkins and Holmes 2003). Dinucleotide biases can be extreme. For example, CpG is present at only 20% of its expected frequency in most vertebrate genomes (Jones et al. 1992) and is depleted in other organisms (Karlin et al. 1994). The most popular explanation for the underrepresentation of CpG in vertebrate genomes is that 60–90% of cytosines in CpG doublets are methylated by cellular methyltransferases (Kress et al. 2001) and methylated cytosine (5-methylcytosine) has a tendency to undergo deamination when unpaired, resulting in the mutation of the cytosine to a thymidine. Methylation can control the expression of particular genes by preventing the binding of transcription factors and modifying chromatin structure and interactions with histones. In vertebrates, methylation is central to genomic imprinting and cell differentiation (Kress et al. 2001). Thus, CpG may be selected for at certain sites and selected against at others. While mutation following methylation probably plays a strong role in the reduced CpG content of many organisms, it is unlikely to be the only factor. In particular, the underrepresentation of CpG is often not accompanied by a correspondingly high level of TpG. Furthermore, similarly low CpG contents are found in vertebrate mitochondrial genomes which are unmethylated (De Amicis and Marchetti 2000).

The dinucleotide TpA is also underrepresented in many genomes (Burge et al. 1992). This is often attributed to the susceptibility of UpA uracils to RNase and that two of the three stop codons in the universal code begin with TpA (Beutler et al. 1989). Additionally, structural factors may have an effect on the frequency of certain dinucleotides. For example, the reduced frequency of TpA may also be explained by its low thermal stability (Breslauer et al. 1986; Beutler et al. 1989), such that TpA-rich sequences in DNA helices may suffer from detrimental levels of unwinding, twisting, and bending (De Amicis and Marchetti 2000).

Eukaryotic DNA viruses can be classified into two or three broad groups based on the size and nature of their genome. The genomes of large double-stranded DNA (dsDNA) viruses are mostly greater than 100 kb in length, while a putative second group, the Adenoviridae, are 28–45 kb in length but biologically similar to the large dsDNA viruses in that they encode their own DNA polymerase and accessory proteins involved in immune response regulation. In contrast, the genomes of small DNA viruses are less than 10 kb in length and comprised of either dsDNA or single-stranded DNA (ssDNA). Large dsDNA viruses encode many different proteins, often exceeding 100, which play intricate roles in virus replication, host-cell regulation, and host immune modulation (Shackelton and Holmes 2004). In con-

trast, small DNA viruses encode fewer than 10 proteins, some of which provide capsid structural functions while the remainder are primarily involved in supplementing or stimulating replication by host machinery, as these viruses do not encode their own polymerases. For the small double-stranded *Polyomaviridae* and *Papillomaviridae*, these supplemental replication proteins typically induce the host cell to enter S phase through a variety of interactions with host cell cycle regulatory components (Cole et al. 2001; Howley et al. 2001). Most of the single-stranded parvoviruses can only replicate in cells that are mitotically active, while the dependo-parvoviruses rely on co-infection of their host cells by a large helper adenovirus or herpesvirus (Muzyczka and Berns 2001). In addition, small DNA viruses are generally unable to modulate either the innate or the adaptive immune response of the host to nearly the extent of the large DNA viruses. This, and the fact that small DNA viruses are far more dependent on cellular proteins and other resources for their propagation, may have resulted in differences in the evolution of biases in base composition and codon usage among large and small viruses.

Herein we conducted an analysis of the codon usage and nucleotide biases among all vertebrate-infecting DNA virus families. Such a comparative analysis is central to revealing the interplay between the genetic drift of neutral mutations and the selective fixation of advantageous ones, can help elucidate the evolutionary dynamics of viruses and their interactions with hosts, and can potentially improve the efficiency and effectiveness of experimental or therapeutic procedures in which viral genome sequences are altered.

Materials and Methods

Sequence Data

Reference sequences (as defined by GenBank; http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/10239.html) were obtained for at least one species within all genera of vertebrate DNA viruses for which such data exist. This resulted in data sets of 41 large dsDNA viruses, 21 small dsDNA viruses, and 15 small ssDNA viruses. The hepadnaviruses were excluded, as they replicate via reverse transcription and are therefore likely to be subject to different evolutionary pressures. Known and putative ORFs were concatenated for total codon analyses. Accession numbers are given in Table 1.

Measuring Codon Usage and Nucleotide Biases

The CodonW package (http://bioweb.pasteur.fr/docs/softgen.html#CODONW) was employed to measure the effective codon usage statistic, Nc (Wright 1990), codon usage frequency, and nucleotide biases. Nc is calculated on a scale of 20 to 61, with a score of 20 representing maximum bias—the use of only one codon for each of the 20 amino acids—and a score of 61 indicating no bias, such that all codons are used equally for each amino acid.

We measured the total G + C (denoted GC) content and the frequency of all dinucleotides within both the concatenated ORFs and the complete genome sequences. Dinucleotide biases were calculated as the observed frequency of the dinucleotide relative to the product of the frequencies of the individual nucleotides (i.e., the expected dinucleotide frequency). For example, $\rho TpA = f(TA)/(f(T)f(A))$. For dinucleotides which do not form a reverse complemented pair on the opposite strand, we symmetrized the measure of ρ with the complementary dinucleotide and labeled this ρ* as outlined by Burge et al. (1992). For example, $\rho^*TpG = 2(f(TG + CA)/[(f(T + A))(f(G + C))]$.

Besides total GC content, we measured the frequency with which either of these nucleotides appears at the synonymous third codon position (GC_{3s}). GC_{3s} was compared to the GC content at the first and second codon positions ($GC_{1,2}$) with the Pearson product-moment correlation coefficient (r). To examine the influence of GC content on codon usage we plotted the relationship between Nc and GC_{3s} for each virus. This was compared to the Nc which would result if GC content were solely responsible for the codon biases (denoted Nc*), calculated as $Nc^* = 2 + GC_{3s} + \{29/[(GC_{3s})^2 + (1 - GC_{3s})^2]\}$ (Wright 1990).

To quantify the effects of natural selection on codon usage we also calculated the Nc' statistic for each data set using the programs SeqCount and ENCprime (Novembre 2002). Nc' is similar to Nc but takes into account background nucleotide composition. This statistic, which also ranges from 20 to 61, measures the deviation of the actual codon usage pattern from the distribution expected from the background nucleotide composition (Novembre 2002).

To determine codon biases across genomes we measured the Nc and GC_{3s}, with the methods described above, of individual genes from each family, subfamily, or distinct genus. To determine CpG biases across large DNA virus genomes we conducted a sliding window analysis of CpG content with a window size of 300 and a step size of 100 using the SWAAP 1.0.2 program (Pride 2000).

Results

Codon Usage Biases

The effective codon usage statistic, Nc, was used to measure codon usage bias across the ORFs of large and small DNA viruses (Table 1). While all viruses exhibit some codon bias, none of these biases were unusually strong and the majority of families include members with both low and high codon biases, following no apparent trend. The mean (and range) for the Nc values of the large and small viruses were 48.3 (34.7–59.6) and 50.5 (35.8–58.0), respectively. However, an examination of genome and gene codon usage tables for a number of viruses suggested that the codon biases which do exist are greatly influenced by GC frequency. For example, in canine parvovirus and human BK virus, which have low Nc values, the most frequently employed synonymous codons are those with the fewest Gs and Cs. As a case in point, of the six synonymous codons coding for arginine, AGA, the only codon with a single G or C, is used 83% and 61% of the time in these two viruses, respectively. In contrast, HSV-2, a virus with a low Nc value which probably results from an overrepresentation of Gs and Cs, employs the two arginine codons with only Cs and Gs, CGC and CGG, 82% of the time.

Table 1. Codon usage bias, as measured by the effective codon usage statistic, not accounting for (Nc) and accounting for (Nc') background nucleotide composition, and nucleotide bias (measured by ORF GC_{3s} and genomic GC content).

	~Genome size (kb)	Nc[a]	Nc′[a]	GC_{3s}^{a}	GC[b]	Accession No.
(A) Large dsDNA viruses						
Adenoviridae	28–45					
Atadenovirus						
Duck adenovirus 1 (A)		55.5	58.5	0.37	0.45	NC_001813
Ovine adenovirus D		45.2	56.3	0.24	0.35	NC_004037
Aviadenovirus: Fowl adenovirus D		54.7	56.7	0.63	0.55	NC_000899
Mastadenovirus:						
Human adenovirus C		50.8	53.6	0.64	0.56	NC_001405
Canine adenovirus		56.4	56.6	0.46	0.49	NC_001734
Murine adenovirus A		59.6	59.4	0.51	0.49	NC_000942
Bovine adenovirus B		45.0	41.3	0.6	0.55	NC_001876
Simian adenovirus 3		50.1	53.0	0.64	0.56	NC_006144
Porcine adenovirusA		39.5	49.8	0.81	0.65	NC_005869
Siadenovirus: Frog adenovirus 1		48.8	56.3	0.3	0.39	NC_002501
Asfarviridae (African swine fever virus)	170–190	57.2	56.7	0.41	0.4	NC_001659
Herpesviridae—Alphaherpesvirinae	130–150					
Bird herpesvirus: Psittacid herpesvirus		48.7	55.8	0.72	0.62	NC_005264
Mardivirus: Gallid herpesvirus 2		56.5	59.0	0.37	0.55	NC_002229
Simplexvirus						
HSV-1/HHV1		39.9	52.3	0.83	0.69	NC_001806
HSV-2/HHV2		37.3	51.5	0.87	0.71	NC_001798
Varicellovirus						
VZV/HHV3		56.1	57.2	0.41	0.47	NC_001348
Equine herpesvirus-1		53.1	56.3	0.64	0.58	NC_001491
Herpesviridae—Betaherpesvirinae	140–240					
Cytomegalovirus						
Cytomegalovirus/HHV5		57.3	56.4	0.48	0.58	NC_001347
Chimpanzee cytomegalovirus		55.8	55.0	0.49	0.63	NC_003521
Muromegalovirus: Murid herpesvirus 2		43.7	51.8	0.78	0.62	NC_002512
Roselovirus: Human herpesvirus 6		58.4	58.2	0.45	0.44	NC_001664
Unclassified: Tupaiid herpesvirus		39	50.9	0.84	0.67	NC_002794
Herpesviridae—unassigned Ictalurid virus	134	48.6	52.7	0.58	0.57	NC_001493
Herpesviridae—Gammaherpesvirinae	110–185					
Lymphocryptovirus						
Callitrichine herpesvirus 3		58.4	58.6	0.45	0.5	NC_004367
EBV/HHV4		50	55.3	0.68	0.61	NC_001345
Rhadinovirus						
Murid herpesvirus 4		55.1	55.0	0.49	0.48	NC_001826
Herpesvirus saimiri		43.1	53.2	0.24	0.36	NC_001350
Iridoviridae	140–383					
Lymphocystivirus: Lymphocystis disease virus		37.4	53.8	0.14	0.29	NC_005902
Ranavirus: Frog virus 3		42.5	44.8	0.73	0.56	NC_005946
Unclassified: Infectious spleen and kidney necrosis virus		51.6	53.9	0.63	0.56	NC_003494
Poxviridae—Chordopoxvirinae	130–375					
Avipoxvirus: Fowlpox virus		44.5	56.9	0.24	0.32	NC_002188
Capripoxvirus: Sheeppox virus		37.8	56.6	0.15	0.26	NC_004002
Leporipoxvirus: Myxoma virus		53.6	52.5	0.52	0.45	NC_001132
Molluscipoxvirus: Molluscum contagiosum virus		38.1	47.6	0.82	0.64	NC_001731
Orthopoxvirus						
Vaccinia virus		48.6	58.4	0.28	0.35	NC_001559
Variola virus		47.3	58.2	0.27	0.34	NC_001611
Monkeypox virus		47.5	58.2	0.27	0.34	NC_003310
Rabbitpox virus		47.9	58.1	0.28	0.35	NC_005858
Parapoxvirus: Orf virus		34.7	44.4	0.88	0.64	NC_005336
Suipoxvirus: Swinepox virus		40.2	56.7	0.18	0.28	NC_003389
Yatapoxvirus: Yaba monkey tumor virus		43.7	55.6	0.26	0.31	NC_005179

(Continued)

Table 1. Continued

	~Genome size (kb)	Nc[a]	Nc′[a]	GC_{3s}^{a}	GC[b]	Accession No.
(B) Small dsDNA viruses						
Papillomaviridae						
Papillomavirus						
Bovine papillomavirus 2	7–8	55.1	56.7	0.42	0.47	NC_001521
Human papillomavirus type 11		50.7	55.8	0.32	0.42	NC_001525
Cottontail papillomavirus		54.5	56.6	0.43	0.47	NC_001541
Chimpanzee papillomavirus type 1		46.9	54.5	0.27	0.39	NC_001838
Deer papillomavirus		57.1	57.7	0.46	0.49	NC_001523
Equus caballus papillomavirus type 1		55	55.9	0.55	0.54	NC_003748
Equinus papillomavirus		54.9	55.8	0.54	0.54	NC_004194
European elk papillomavirus		57.9	58.4	0.46	0.48	NC_001524
Human papillomavirus type 43		49.8	55.3	0.31	0.41	NC_005349
Multimammate rat papillomavirus		56.1	56.4	0.47	0.51	NC_001605
Canine oral papillomavirus		52	56.6	0.36	0.43	NC_001619
Monkey B-lymphotropic papovavirus		45.4	50.6	0.32	0.42	NC_001536
Polyomaviridae	5					
Polyomavirus						
Human JC virus		44.7	50.1	0.3	0.41	NC_001699
Human BK virus		43.2	49.5	0.28	0.41	NC_001538
Simian virus 40		44.2	48.6	0.31	0.42	NC_001669
Budgerigar fledgling polyomavirus		58	58.2	0.45	0.5	NC_004764
Murine polyomavirus		52.5	53.3	0.42	0.48	NC_001515
African green monkey polyomavirus		46.6	51.4	0.32	0.42	NC_004763
Goose hemorrhagic polyomavirus		52.7	53.2	0.43	0.42	NC_004800
Hamster polyomavirus		47.8	52.5	0.33	0.43	NC_001663
Bovine polyomavirus		48.4	53.0	0.33	0.42	NC_001442
(C) Small ssDNA viruses						
Anellovirus: Human TT virus	4	51.9	55.7	0.43	0.5	NC_002076
Circoviridae	2					
Circovirus						
Porcine circovirus 2		54.4	54.6	0.5	0.5	NC_005148
Porcine circovirus 1		54.3	50.9	0.52	0.51	NC_006266
Bovine circovirus		58	56.6	0.5	0.5	NC_002068
Canary circovirus		55.7	56.2	0.54	0.54	NC_003410
Gyrovirus: Chicken anemia virus		55.4	58.5	0.58	0.57	NC_001427
Parvoviridae—Parvovirinae	4–6					
Dependovirus						
Adeno-associated virus 2		49.2	51.2	0.67	0.55	NC_001401
Adeno-associated virus 8		45	50.3	0.72	0.58	NC_006261
Avian adeno-associated virus		53	54.7	0.63	0.54	NC_006263
Bovine adeno-associated virus		49.3	50.3	0.62	0.55	NC_005889
Amdovirus: Aleutian mink disease virus		43.9	51.0	0.33	0.41	NC_001662
Bocavirus: Bovine parvovirus		49.4	52.8	0.51	0.50	NC_001540
Erythrovirus: B19		46.9	50.9	0.31	0.45	NC_000883
Parvovirus						
Canine parvovirus		35.8	47.5	0.16	0.38	NC_001539
Mouse minute virus		43.7	48.6	0.32	0.43	NC_001510

[a]Concatenated ORFs.
[b]Whole genome.

Nucleotide Biases

To determine whether these codon biases are primarily caused by mutation pressure on overall base composition or natural selection for particular triplets, we first explored the extent of underlying nucleotide bias in all viral genomes. We began by examining the overall GC content at the genomic level and in the concatenated ORFs, as well as the GC_{3s} content of the latter (Table 1). Although values varied greatly, the majority of large viruses, including the adenoviruses, alphaherpesviruses, and betaherpesviruses, have GC_{3s} contents above 0.50, averaging 0.52, 0.64, and 0.61, respectively. However, specific families of large DNA viruses, namely, the Poxviridae, the Asfarviridae, and some iridoviruses,

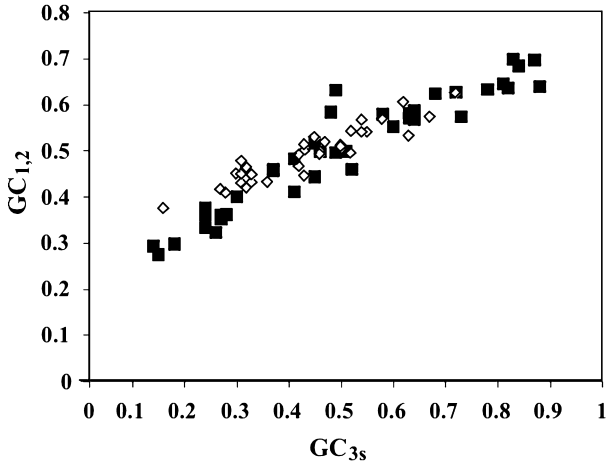

Fig. 1. Correlation between GC content at the synonymous third codon position (GC_{3s}) and GC content at the nonsynonymous first/second positions ($GC_{1,2}$) of DNA viruses. Filled squares and open diamonds represent the individual large DNA viruses and small DNA viruses used in this study, respectively.

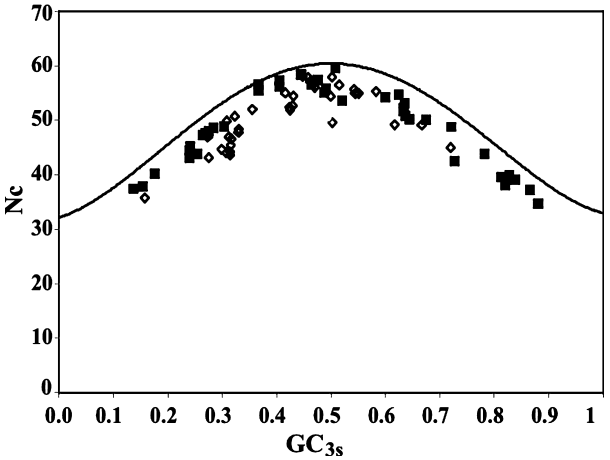

Fig. 2. Correlation between the synonymous third codon position (GC_{3s}) and the effective codon usage statistic (Nc) for DNA viruses. Filled squares and open diamonds represent the individual large DNA viruses and small DNA viruses, respectively. As shown, these Nc values form a curve on, or slightly below, the line representing the expected Nc values (Nc*) which would result if GC composition were the only factor influencing codon usage bias.

have low GC_{3s} contents. Interestingly, these are the only DNA viruses which replicate completely (Poxviridae) or partially in the cytoplasm of the cell rather than entirely in the nucleus (Moss 2001; Rojo et al. 1999; Williams 1996). Almost all small DNA viruses showed low GC_{3s} values. For the Papillomaviridae, Polyomaviridae, anellovirus, Circoviridae, and Parvoviridae (excluding the dependovirus genera of the Parvoviridae, which do not show this trend), the average GC_{3s} values were 0.41, 0.35, 0.43, 0.53, and 0.33, respectively. For all viruses, GC_{3s} values were close to their genomic GC values. There were no obvious trends in single nucleotide strand biases across virus families.

Mutation Pressure Versus Codon Selection

To determine the relative effects of mutation pressure versus natural selection on codon composition, we examined the relationship between GC_{3s} and GC at the first and second codon positions ($GC_{1,2}$). As shown in Fig. 1, GC_{3s} and $GC_{1,2}$ were significantly correlated ($r = 0.95$ for the large DNA viruses and $r = 0.94$ for the small DNA viruses; $p < 0.001$). Although this test does not take into account the phylogenetic relationships among the taxa studied, it does show that observed patterns of base composition are present at all codon positions. This suggests that they are most likely the result of mutation pressure, as natural selection would be expected to act differently on different codon positions.

The notion that codon bias is primarily governed by an underlying biased mutation pressure was further supported by examining the correlation between the GC_{3s} and Nc of the coding sequences. We plotted GC_{3s} against Nc for each virus and compared this to the expected Nc (Nc*) that would result if GC content were the sole determinant of codon usage. Remarkably, the actual Nc values for most viruses were on, or just below, the Nc* curve (Fig. 2). This implies that codon bias is mainly explained by uneven base composition and, hence, by mutation pressure rather than natural selection on codon choice. However, the fact that the majority of the actual Nc values were slightly lower than Nc* indicates that there are other factors, with less of an effect, that also influence the codon bias.

Additional evidence that codon biases are predominantly influenced by mutation pressure was obtained by determining the Nc' statistic for each virus. Unlike Nc, Nc' takes into account the background nucleotide composition and should therefore reflect the degree of codon usage bias due to factors other than nucleotide composition. As expected, Nc' was greater than Nc in the vast majority of viruses and the difference between Nc and Nc' was greatest in those viruses which have GC contents that depart significantly from 50% and minimal in those viruses with GC contents near 50% (Table 1). The graphic relationship between GC content and Nc'–Nc is a V-shaped curve centered at a GC content of 0.5—approximated by $y = C(|x - 0.5|)$ (Fig. 3).

Dinucleotide Biases

Because there have been reports of CpG underrepresentation in RNA and small DNA viruses (Karlin et al. 1994) and because dinucleotide biases can affect codon bias, we next determined the dinucleotide biases for each virus. We measured the actual frequency of each pair of nucleotides relative to the

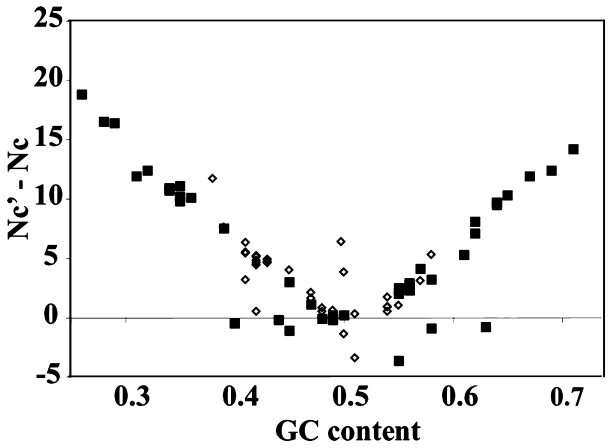

Fig. 3. Graphical representation of the relationship between GC content and Nc' – Nc. The difference between Nc' and Nc depicted here should reflect the amount of bias due to background base composition. As shown, this value is greatest in viruses with extreme GC contents. Filled squares and open diamonds represent the individual large DNA viruses and small DNA viruses, respectively.

expected frequency (i.e., the product of the frequencies of the individual nucleotides). As previously observed by Karlin et al. (1994), while most large DNA viruses showed no bias against CpG, small DNA viruses were extremely biased against this dinucleotide (Table 2, Fig. 4). This difference in CpG content was highly significant (Student's *t*-test: $p = 2.3^{-15}$ when outliers were excluded [see below] and $p = 2.2^{-11}$ when outliers were included) and indicates that in small DNA viruses there is additional mutation pressure away from this doublet which may also impact codon usage. The CpG content of the small papillomaviruses, polyomaviruses, anellovirus, circoviruses, and parvoviruses (excluding the dependoviruses) averaged 47, 22, 67, 62, and 43% of their expected values, respectively. The outliers, the dependoviruses, averaged 82%, which is significantly different from the other small DNA viruses ($p = 0.005$). In contrast, the large viruses—the adenoviruses, asfarvirus, alphaherpesviruses, betaherpesviruses, unassigned herpesvirus, and poxviruses—had average CpG contents of 78, 87, 98, 115, 108, and 112% of their expected values. However, among the large DNA viruses there are two distinct exceptions (significant at $p = 0.0003$) to the general correlation between genome size and CpG content: the Gammaherpesvirinae and the Iridoviridae, which have CpG contents 49 and 70% of their expected values, respectively.

In those viruses with CpG depletion there were, generally, also elevated levels of TpG (measured as discussed above in order to symmetrize the value over both strands), which would result from a cytosine-to-thymidine mutation. However, the overrepresentation of this pair was slight compared to the underrepresentation of CpG. The genomic CpG and TpG contents were both approximately equivalent to the ORF contents. Finally, and as seen in most organisms, the TpA doublet had a reduced frequency in both large and small DNA viruses, although this underrepresentation was small compared to the level of CpG suppression in the small viral genomes (Table 2). With the exception of those mentioned above, all other dinucleotides were present at approximately the expected frequency.

Comparisons Along Viral Genomes

To determine the extent to which codon biases vary among genes, we selected one virus from each family, subfamily, or distinct genus and examined every gene separately. Because correct gene prediction is difficult for large viruses and results in some annotations that are only putative, we limited our analysis to those genes with assigned functions in the NCBI database. Strikingly, viral genes located in different genomic regions do not differ dramatically in the biases they display (see Supplementary Table).

Furthermore, because large genomes can have very different nucleotide compositions in different regions, we used a sliding window method to examine CpG content along the genomes of the large viruses. While small windows varied in their CpG contents, we found no systematic diversion from the average viral bias in any genomic region (see Supplementary Figure).

Discussion

This study revealed a number of trends in the nucleotide and codon composition among families of vertebrate-infecting DNA viruses. First, the strong correlation between codon usage bias and GC composition indicates that codon usage bias in DNA viruses is primarily explained by overall nucleotide content. In addition, not only are the GC frequencies of each virus similar at nonsynonymous and synonymous codon positions, but these frequencies appear in genes with different genomic positions and functions. These observations suggest that genome-wide mutational pressure is the most important factor shaping patterns of codon usage bias in DNA viruses, rather than natural selection for specific codons. Although nucleotide biases can vary even among closely related viruses, those viruses with similar genomes and life-cycle characteristics most frequently show similar CpG frequencies relative to their GC content. Such generalities, especially the marked CpG deficiency in the genomes of small DNA viruses, along with the relatively high levels of this dinucleotide in the large DNA viruses, point to common evolutionary pressures faced by similar viruses. These obser-

Table 2. Dinucleotide biases of the viral genomes, measured with ρ, the observed frequency of the dinucleotide relative to the expected frequency

	ρCpG	ρ*TpG	ρTpA
(A) Large dsDNA viruses			
Adenoviridae			
Atadenovirus			
Duck adenovirus 1 (A)	0.74	1.15	0.89
Ovine adenovirus D	0.51	1.11	0.79
Aviadenovirus: Fowl adenovirus D	1.13	0.88	0.92
Mastadenovirus			
Human adenovirus C	0.86	1.13	0.83
Canine adenovirus	0.54	1.16	0.84
Murine adenovirus A	0.78	1.13	0.78
Bovine adenovirus B	0.79	1.09	0.82
Simian adenovirus 3	0.89	1.08	0.73
Porcine adenovirus A	0.8	1.18	0.64
Siadenovirus: Frog adenovirus 1	0.53	1.18	0.75
Asfarviridae (African swine fever virus)	0.87	1.02	0.92
Herpesviridae—Alphaherpesvirinae			
Bird herpesvirus: Psittacid herpesvirus	1.17	0.92	0.94
Mardivirus: Gallid herpesvirus 2	0.65	1.05	1.1
Simplexvirus			
HSV-1/HHV1	0.98	1.00	0.83
HSV-2/HHV2	1.04	0.96	0.76
Varicellovirus			
VZV/HHV3	1.07	1.00	0.98
Equine herpesvirus-1	0.96	1.02	0.96
Herpesviridae—Betaherpesvirinae			
Cytomegalovirus			
Cytomegalovirus/HHV5	1.15	1.05	0.85
Chimpanzee cytomegalovirus	1.08	1.04	0.8
Muromegalovirus: Murid herpesvirus 2	1.21	0.82	0.8
Roselovirus: Human herpesvirus 6	1.08	1.02	0.82
Unclassified: Tupaiid herpesvirus	1.25	0.86	0.79
Herpesviridae—unassigned Ictalurid virus	1.08	0.99	0.72
Herpesviridae—Gammaherpesvirinae			
Lymphocryptovirus			
Callitrichine herpesvirus 3	0.66	1.14	0.85
EBV/HHV4	0.58	1.14	0.79
Rhadinovirus			
Murid herpesvirus 4	0.41	1.24	0.77
Herpesvirus saimiri	0.32	1.21	0.87
Iridoviridae			
Lymphocystivirus: Lymphocystis disease virus	0.59	1.00	0.98
Ranavirus: Frog virus 3	0.69	1.06	0.79
Unclassified: Infectious spleen and kidney necrosis virus	0.84	1.52	0.82
Poxviridae—Chordopoxvirinae			
Avipoxvirus: Fowlpox virus	1.13	0.83	1.15
Capripoxvirus: Sheeppox virus	0.84	1.06	0.92
Leporipoxvirus: Myxoma virus	1.61	0.76	1.09
Molluscipoxvirus: Molluscum contagiosum virus	1.2	1.15	0.64
Orthopoxvirus			
Vaccinia virus	1.04	0.98	1.01
Variola virus	0.98	1.00	1.03
Monkeypox virus	0.95	1.01	1.03
Rabbitpox virus	1.04	0.98	1.02
Parapoxvirus: Orf virus	1.25	1.02	0.72
Suipoxvirus: Swinepox virus	1.09	0.99	1.08
Yatapoxvirus: Yaba monkey tumor virus	1.23	0.91	0.96
(B) Small dsDNA viruses			
Papillomaviridae			
Papillomavirus			
Bovine papillomavirus 2	0.42	1.26	0.82

(Continued)

Table 2. Continued

	ρCpG	ρ*TpG	ρTpA
Human papillomavirus type 11	0.43	1.34	1.05
Cottontail papillomavirus	0.53	1.27	0.74
Chimpanzee papillomavirus type 1	0.39	1.35	1.0
Deer papillomavirus	0.46	1.25	0.8
Equus caballus papillomavirus type 1	0.57	1.16	0.9
Equinus papillomavirus	0.57	1.14	0.92
European e lk papillomavirus	0.56	1.23	0.75
Human papillomavirus type 43	0.42	1.30	1.04
Multimammate rat papillomavirus	0.54	1.17	0.95
Canine oral papillomavirus	0.41	1.20	0.81
Monkey B-lymphotropic papovavirus	0.16	1.12	0.82
Polyomaviridae			
Polyomavirus			
Human JC virus	0.07	1.24	0.76
Human BK virus	0.05	1.14	0.88
Simian virus 40	0.12	1.30	0.73
Budgerigar fledgling polyomavirus	0.78	1.06	1.08
Murine polyomavirus	0.29	1.22	0.83
African green monkey polyomavirus	0.18	1.12	0.84
Goose hemorrhagic polyomavirus	0.13	1.20	0.82
Hamster polyomavirus	0.16	1.22	0.81
Bovine polyomavirus	0.18	1.16	0.84
(C) Small ssDNA viruses			
Anellovirus: Human TT virus	0.67	1.06	1.0
Circoviridae			
Circovirus			
Porcine circovirus 2	0.55	1.10	0.84
Porcine circovirus 1	0.49	1.02	1.0
Bovine circovirus	0.52	1.10	0.88
Canary circovirus	0.64	1.02	0.76
Gyrovirus: Chicken anemia virus	0.87	1.06	0.94
Parvoviridae—Parvovirinae			
Dependovirus			
Adeno-associated virus 2	0.77	1.16	0.57
Adeno-associated virus 8	0.78	1.18	0.56
Avian adeno-associated virus	1.03	1.01	0.64
Bovine adeno-associated virus	0.68	1.13	0.56
Amdovirus: Aleutian mink disease virus	0.31	1.16	1.07
Bocavirus: Bovine parvovirus	0.71	1.25	0.78
Erythrovirus: B19	0.42	1.20	0.97
Parvovirus			
Canine parvovirus	0.37	1.18	0.91
Mouse minute virus	0.32	1.30	0.75

Note. In the case where the dinucleotide is not symmetrical, the measure is symmetrized over both strands (ρ*).

vations notwithstanding, it is likely that experimental analyses of DNA viruses, such as artificially altering codon usage or dinucleotide frequencies, are required to fully understand the complex mechanistic basis of these different genomic signatures.

Nucleotide Compositions

It has long been known that viruses may take on a much wider range of GC frequencies than other organisms (Wyatt 1952; Bronson and Anderson 1994). Even viruses within the same family, which have similar replication and life-cycle strategies, can show very different GC contents and hence large differences in codon biases (Schachtel et al. 1991). For example, HSV-2 and VZV (HSV-3), closely related alphaherpesviruses, which infect humans, persist latently in the nervous system of the host, and reactivate to cause secondary infections (but which have different routes of infection and cell tropism), show GC contents of 71 and 47% and Nc values of 37 and 56, respectively. Yet despite these extreme differences in nucleotide and codon bias, the viruses exhibit almost-identical frequencies of CpG doublets in proportion to their GC contents. Again, this supports the idea that there are common mechanisms which determine CpG frequencies for viruses with similar life cycles.

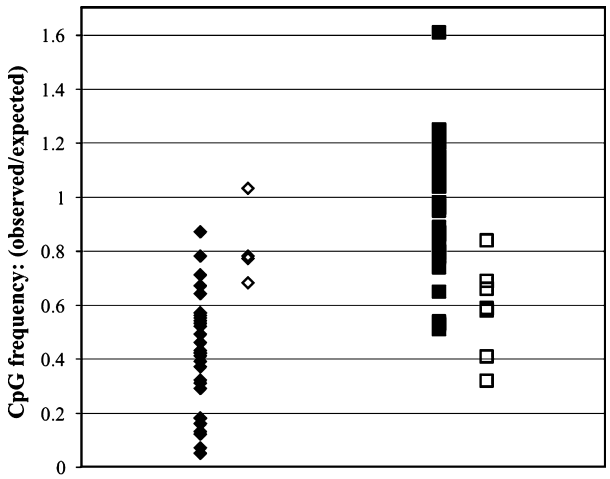

Fig. 4. Genomic ρCpG values (observed CpG frequency/expected CpG frequency) of large and small DNA viruses. Small autonomous viruses are depicted by filled diamonds and small dependoviruses by empty diamonds, while most large viruses are depicted by filled squares and the iridoviruses and gammaherpesviruses by open squares.

Another important observation in this context is that many poxviruses, many iridoviruses, and African swine fever virus have low GC contents. From the perspective of the mutation pressure hypothesis this could be viewed as resulting from their cytoplasmic site of replication, particularly if the nucleotide composition in this location differs greatly from that of the cell nucleus (Moss 2001; Moyer and Henderson 1985; Williams 1996; Rojo et al. 1999). Accordingly, the low GC content in most RNA viruses may similarly reflect both their cytoplasmic site of replication and the fact that they, like small DNA viruses, do not encode enzymes that alter dNTP concentrations.

Comparisons of Host-Viral Genome Compositions

Given the role played by mutation pressure in shaping codon and nucleotide biases in both animals and viruses, and because small DNA viruses are replicated, and possibly repaired, by cellular machinery, whereas large DNA viruses generally encode much of their own replicative machinery, it might be expected that the former will show biases more similar to those of host cellular DNA than the latter. Clearly, viral-host genome comparisons are critical in addressing this question. However, because vertebrate genomes are so large and their base compositions differ dramatically depending on the region of the genome examined, it is difficult to make generalized statements regarding host nucleotide biases. Moreover, unlike the situation in bacteria, there is no strong evidence of a link between codon usage bias and gene expression levels in mammals, as expected under tRNA-mediated selection (Sharp and Matassi 1994). This was confirmed during the initial analysis of the human genome sequence in which codon usage bias was found to be largely determined by genome location and hence local mutation pressure (Lander et al. 2001). Consequently, studies that attempt to compare viral codon usage with host codon usage or tRNA availability are likely to be unreliable.

CpG Frequency

Considerable data have been compiled on CpG contents within different regions of the human genome which are useful in comparing the base composition of human viruses with those of their hosts. CpG suppression in most small human DNA viruses falls within the range of that in the human genome, in which ρCpG values of 50-kb stretches range from 0.12 to 0.45 and chromosomal averages range from 0.18 to 0.31 (Karlin and Mrázek 1997; Gentles and Karlin 2001). However, unlike the situation in their hosts, CpG is not suppressed in most large human DNA viruses, where the frequency of this dinucleotide is almost four times as high as for human DNA. At first sight this suggests that CpG bias in the small DNA viruses could be the result of biases intrinsic to the host replicative/repair machinery/processes which act on small viral genomes.

CpG depletion in vertebrates is largely attributed to the methylation/deamination/mutation of cytosines in these doublets. This is partially supported by the overrepresentation of TpG in the human genome. However, this overrepresentation is slight in comparison to the great underrepresentation of CpG, indicating that methylation cannot be the only cause of CpG depletion in humans (Gentles and Karlin 2001). While small DNA viruses also have elevated levels of TpG, this elevation is even less than that in the human genome, which makes methylation/deamination an even more questionable cause of CpG suppression in these viruses. Furthermore, available evidence indicates that there may be little or no methylation in many viral genomes when they are actively replicated or packaged (Acken et al. 1979; Karlin and Burge 1995; Lundberg et al. 2003; Kämmer and Doerfler 1995).

While the common biases in human and small DNA virus genomes may still be due to shared replication machinery within the nucleus, a few factors point away from this explanation. First, it does not explain the human polyomaviruses which have ρCpG levels of 0.05 and 0.07—far below human levels. Second, small DNA viruses may not employ the complete set of host replicative machinery, as at least some of these viruses have mutation rates orders of magnitude greater than those of the host (Truyen et al. 1995; Shackelton et al. 2005). Finally, it does not explain why RNA viruses, which encode their own polymerases and replicate in the cytoplasm,

should also have many of the same dinucleotide signatures (Karlin et al. 1994).

It is therefore likely that additional factors influence the suppression of CpG in small DNA viruses. The deficiency may be related to the immunostimulatory properties of unmethylated CpGs, which are recognized by the host's innate immune system as a pathogen signature (Krieg 2003). These sequences bind and activate Toll-like receptor 9 (TLR9) on neutrophils, dendritic cells, and macrophages, inducing a rapid immune response (Lund et al. 2003; Wagner 2004). While a high CpG content may be detrimental to small DNA (and RNA) viruses, large ones may not be similarly affected because they encode a range of proteins that interfere with cellular pathogen-pattern recognition. For example, vaccinia poxvirus encodes agonists of TLRs (Harte et al. 2003). It has even been suggested that some complex viruses have evolved mechanisms to activate TLRs to induce the accompanying activation of specific cells (Rassa and Ross 2003). As long as the virus can modulate the antiviral response which accompanies TLR activation, it may induce cellular proliferation for its own replicative advantage.

In contrast to most large viruses, gammaherpesviruses and iridoviruses show a CpG deficiency. This may be because these viruses methylate their genomes at specific times in their life cycle (Ambinder et al. 1999; Willis and Granoff 1980; Wagner et al. 1985), perhaps to protect against immune recognition (Tao and Robertson 2003) and control patterns of transcription. As with vertebrate genomes, methylated viral genomes would face a high chance of mutation at CpGs, which would result in a reduction of this dinucleotide (Ambinder et al. 1999).

The most obvious difference between the CpG-suppressed small DNA viruses and the large DNA viruses is gene number. While the compact genomes of the papillomaviruses, polyomaviruses, anellovirus, circoviruses, and parvoviruses encode only a small number of replication and structural genes, most large viruses encode over 100 genes which, among other things, may allow them to evade adaptive or innate immune detection, suppress antiviral immune reactions, regulate and dominate cellular metabolic machinery, and produce, and compensate for, cellular depletions in metabolic resources (Shackelton and Holmes 2004). Small DNA viruses, which do not encode such an extended array of proteins capable of manipulating host cell metabolism or reducing/delaying host immune responses, may be under strong selection to propagate quickly in a cell with limited resources before host responses resolve the infection. In this case CpG may be selected against not only because of its immunostimulatory effects, but because it may extend the amount of time necessary for viral replication and transcription in the infected cell. Specifically, CpG has the highest stacking energy of any dinucleotide, thereby requiring the greatest amount of free energy to disrupt a double helix (Karlin and Burge 1995; Breslauer et al. 1986).

Finally, that CpG introduces structural abnormalities into the DNA helix (El Antri et al. 1993a, b; Grzeskowiak et al. 1991) may also affect the dinucleotide composition of small DNA viruses, which have few epigenetic protection mechanisms against helical distortions, and the DNA-binding molecules they attract, which may interfere with genome recognition or packaging.

Dependoviruses

Dependoviruses are small viruses of the *Parvoviridae*, so called because they require a helper virus, most often an adenovirus, or in some cases a herpesvirus, for a productive infection. Curiously, these viruses do not show the genomic biases common to other similarly sized DNA viruses, including the closely related autonomous parvoviruses. Instead of CpG depletion, the genomic signature of dependoviruses resembles that of adenoviruses and other large dsDNA viruses. Unlike other parvoviruses, these viruses integrate into the host genome when the helper virus is absent (Muzyczka and Berns 2001). Although the possibly prolonged integrated stage in their life cycle may affect their base composition, helper virus coinfection may also play a role. By replicating in cells alongside these large viruses, dependoviruses may "benefit" from the adenovirus or herpesvirus modulation of the host immune system, control of the cell cycle, and induction or production of supplementary resources. The additional time and resources available for replication and expression, and the suppression of viral detection and immune responses, could result in a dependovirus composition similar to that of its helper virus.

Acknowledgments. This work was completed under a Howard Hughes Medical Institute Fellowship to L.A.S. and NIH Grant R01AI028385 to C.R.P.

References

Acken UV, Simon D, Grunert F, Döring H-P, Kröger H (1979) Methylation of viral DNA in vivo and in vitro. Virology 99:152–157

Ambinder RF, Robertson KD, Tao Q (1999) DNA methylation and the Epstein-Barr virus. Semin Cancer Biol 9:369–375

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Beutler E, Gelbart T, Han J, Koziol JA, 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

Breslauer KJ, Frank R, Blocker H, Marky LA (1986) Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci USA 83:3746–3750

Bronson EC, Anderson JN (1994) Nucleotide composition as a driving force in the evolution of retroviruses. J Mol Evol 38:506–532

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

Chamary J-V, Hurst LD (2004) Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents, evidence for selectively driven codon usage. Mol Biol Evol 21:1014–1023

Cole CN, Conzen SD (2001) *Polyomaviridae*: the viruses and their replication. In: Knipe DM, Howley PM (eds) Fundamental virology, vol 4. Lippincott Williams and Wilkins, Philadelphia, PA, pp 985–1018

De Amicis F, Marchetti S (2000) Intercodon dinucleotides affect codon choice in plant genes. Nucleic Acids Res 28:3339–3345

Duan J, Antezana MA (2003) Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. J Mol Evol 57:694–701

El Antri S, Bittoun P, Mauffret O, Monnot M, Lescot E, Convert O, Fermandjian S (1993a) Effect of distortions in the phosphate backbone conformation of six related octanucleotide duplexes on CD and 31P NMR spectra. Biochemistry 32:7079–7088

El Antri S, Mauffret O, Monnot M, Lescot E, Convert O Fermandjian S (1993b) Structural deviations at CpG provide a plausible explanation for the high frequency of mutation at this site, Phosphorus nuclear magnetic resonance and circular dichroism studies. J Mol Biol 230:373–378

Gentles AJ, Karlin S (2001) Genome-scale compositional comparisons in eukaryotes. Genome Res 11:540–546

Gouy M, Gautier C (1982) Codon usage in bacteria, correlation with gene expressivity. Nucleic Acids Res 10:7055–7047

Grantham R, Gautier C, Guoy M, Mercier R, Pave A (1980) Codon catalogue usage and the genome hypothesis. Nucleic Acids Res 8:49–62

Grantham R, Perrin P, Mouchiroud D (1986) Patterns in codon usage of different kinds of species. Oxford Surv Evol Biol 3:48–81

Grzeskowiak K, Yanagi K, Privé GG, Dickerson RE (1991) The structure of B-helical C-G-A-T-C-G-A-T-C-G and comparison with C-C-A-A-C-G-T-T-G-GJ. Biol Chem 266:8861–8883

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

Harte MT, Haga IR, Maloney G, Gray P, Reading PC, Bartlett NW, Smith GL, Bowie A, O'Neill AJ (2003) The poxvirus protein A52R targets toll-like receptor signalling complexes to suppress host defense. J Exp Med 197:343–351

Howley PM, Lowy DR (2001) Papillomaviruses and their replication. In: Knipe DM, Howley PM (eds) Fundamental virology, vol 4. Lippincott Williams and Wilkins, Philadelphia, PA, pp 1019–1051

Jenkins GM, Holmes EC (2003) The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 92:1–7

Jones PA, Rideout WMIII, Shen J-C, Spruck CH, Tsai YC (1992) Methylation, mutation and cancer. Bioessays 14:33–36

Kämmer C, Doerfler W (1995) Genomic sequencing reveals absence of DNA methylation in the major late promoter of adenovirus type 2 DNA in the virion and in productively infected cells. FEBS Lett 362:301–305

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

Karlin S, Mrázek J (1997) Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci USA 94:1027–10232

Karlin S, Doerfler W, Cardon LR (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

Kress C, Thomassin H, Grange T (2001) Local DNA methylation in vertebrates, how could it be performed and targeted? FEBS Lett 494:135–140

Krieg AM (2003) CpG DNA, Trigger of sepsis, mediator of protection, or both? Scand J Infect Dis 35:653–659

Lander ES, Linton LM, Birren B et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lund J, Sato A, Medzhitov R, Iwasaki A (2003) Toll-like receptor 9-mediated recognition of Herpes simplex virus-2 by plasmacytoid dendritic cells. J Exp Med 198:513–520

Lundberg P, Welander P, Han X, Cantin E (2003) Herpes simplex virus type 1 DNA is immunostimulatory in vitro and in vivo. J Virol 77:11158–11169

Moss B (2001) *Poxviridae*: The viruses and their replication. In: Knipe D, Howley P (eds) Fundamental virology, vol 4. Lippincott Williams and Wilkins, Philadelphia, PA, pp 1249–1283

Moyer JD, Henderson JF (1985) Compartmentation of intracellular nucleotides in mammalian cells. CRC Crit Rev Biochem 19:45–61

Muzyczka N, Berns KI (2001) *Parvoviridae*: the viruses and their replication. In: Knipe DM, Howley PM (eds) Fundamental virology, vol 4. Lippincott Williams and Wilkins, Philadelphia, PA, pp 1089–1121

Novembre JA (2002) Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol 19:1390–1394

Oresic M, Shalloway D (1998) Specific correlations between relative synonymous codon usage and protein secondary structure. J Mol Biol 281:31–48

Powell JR, Moriyama EN (1997) Evolution of codon usage bias in Drosophila. Proc Natl Acad Sci USA 94:7784–7790

Pride DT (2000) SWAAP Version 1.0.0—Sliding windows alignment analysis program: a tool for analyzing patterns of substitutions and similarity in multiple alignments. Distributed by the author

Rassa J, Ross SR (2003) Viruses and toll-like receptors. Microbes Infect 5:961–968

Rojo G, García-Beato R, Viñuela E, Sala MA, Salas J (1999) Replication of African swine fever virus DNA in infected cells. Virology 257:542–536

Schachtel GA, Bucher P, Mocarski ES, Blaisdell BE, Karlin S (1991) Evidence for selective evolution in codon usage in conserved amino acid segments of human alphaherpesvirus proteins. J Mol Evol 33:483–494

Shackelton LA, Holmes EC (2004) The evolution of large DNA viruses, combining genomic information of viruses and their hosts. Trends Microbiol 12:458–465

Shackelton LA, Parrish CR, Truyen U, Holmes EC (2005) High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc Natl Acad Sci USA 102:379–384

Sharp PM, Matassi G (1994) Codon usage and genome evolution. Curr Opin Genet Dev 4:851–860

Sharp PM, Tuohy TM, Mosurski KR (1986) Codon usage in yeast, cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14:1525–5143

Sharp PM, Stenico M, Peden JF, Lloyd AT (1993) Codon usage, mutational bias, translational selection, or both? Biochem Soc Trans 21:835–841

Smith NGC, Eyre-Walker A (2001) Synonymous codon bias is not caused by mutation bias in G + C-rich genes in humans. Mol Biol Evol 18:982–986

Stenico M, Lloyd AT, Sharp PM (1994) Codon usage in *Caenorhabditis elegans*, delineation of translational selection and mutational biases. Nucleic Acids Res 22:2437–2446

Strauss EG, Strauss JH, Levine AJ (1996) Virus evolution. In: Fields BN, Knipe DM, Howley PM (eds) Virology. Lippincott-Raven, Philadelphia, PA, pp 153–171

Sueoka N (1961) Compositional correlation between deoxyribonucleic acid and protein. Cold Spring Harbor Symp Quant Biol 26:35–43

Tao Q, Robertson KD (2003) Stealth technology, how Epstein–Barr virus utilizes DNA methylation to cloak itself from immune detection. Clin Immunol 109:53–63

Truyen U, Gruenberg A, Chang SW, Obermaier B, Veijalainen P, Parrish CR (1995) Evolution of the feline-subgroup parvoviruses and the control of canine host range in vivo. J Virol 69: 4702–4710

Wagner H (2004) The immunobiology of the TLR9 subfamily. Trends Immunol 25:381–386

Wagner H, Simon D, Werner E, Gelderblom H, Darai C, Flügel RM (1985) Methylation pattern of fish lymphocystis disease virus DNA. J Virol 53:1005–1007

Williams T (1996) The iridoviruses. Adv Virus Res 46:345–412

Willis DB, Granoff A (1980) Frog virus 3 DNA is heavily methylated at CpG sequences. Virology 107:250–257

Wright F (1990) The ''effective number of codons'' used in a gene. Gene 87:23–29

Wyatt GR (1952) The nucleic acids of some insect viruses. J Gen Physiol 36:201–205

Xia X (1996) Maximizing transcription efficiency causes codon usage bias. Genetics 144:1309–1320

Zama M (1990) Codon usage and secondary structure of mRNA. Nucleic Acids Symp Ser 22:93–94

Zhao K-N, Liu WJ, Frazer IH (2003) Codon usage bias and A + T content variation in human papillomavirus geomes. Virus Res 98:95–104