JOURNAL OF VIROLOGY, May 1994, p. 2889–2897
0022-538X/94/$04.00+0

Vol. 68, No. 5

Why Is CpG Suppressed in the Genomes of Virtually All Small Eukaryotic Viruses but Not in Those of Large Eukaryotic Viruses?

S. KARLIN,[1]* W. DOERFLER,[2] AND L. R. CARDON[1]

Department of Mathematics, Stanford University, Stanford, California 94305,[1] and Institute for Genetics, University of Cologne, Cologne, Germany[2]

Received 11 December 1993/Accepted 26 January 1994

Dinucleotide over- and underrepresentation is evaluated in all available completely sequenced DNA or RNA viral genomes, ranging in size from 3 to 250 kb (available RNA viruses fall into the small-virus category). The dinucleotide CpG is statistically underrepresented (suppressed) in all but four of the small viruses (more than 75 with lengths of <30 kb) but has normal relative abundances in most large viruses (≥30 kb). Most retrotransposons in eukaryotic species also show low CpG relative abundances. Interpretations, especially in some cases of DNA viruses or viruses with a DNA intermediate, might relate to methylation effects and modes of viral integration and excision. Other possible contributing factors relate to dinucleotide stacking energies, special mutation mechanisms, and evolutionary events.

Vertebrate genomic DNA is pervasively CpG suppressed. The traditional explanation for this centers on methylation of CpG dinucleotides at position 5 of the cytosine base, which through deamination of 5-methylcytosine (possibly enzymatically mediated) and failure to repair the mismatch mutates to TpG/CpA. At least 60% of CpG in some sequences in vertebrate DNA is methylated (2, 51). The CpG dinucleotide relative abundance is normal in almost all invertebrate and fungal species (e.g., *Drosophila melanogaster*, *C. elegans*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Neurospora crassa*) and most common bacteria (8), presumably because of the absence of the standard methylase in these species. Exceptions include the archebacterium *Methanobacterium thermoautotrophicum*, the primitive *Thermus thermophilus*, and the parasitic *Mycoplasma capricolum*. Our principal observations are as follows.

(i) All small vertebrate viral genomes (including more than 75 completely sequenced genomes of <30 kb in length), apart from those of four togaviruses, are significantly CpG suppressed (see Tables 1 and 4).

(ii) In all large or intermediate-size viral genomes (≥30 kb), apart from those of the gamma herpesviruses, CpG relative abundances are in the normal range (see Table 1).

(iii) In retrotransposons (≥5 kb in length) of eukaryotic species, CpG dinucleotides are often of low relative abundance (see Table 5).

Some interpretations and hypotheses related to methylation patterns, dinucleotide stacking energies, preferential mutations, and evolutionary processes are considered.

METHODS

A common assessment of dinucleotide bias in a single sequence is via the odds-ratio measure, $\rho_{XY} = f_{XY}/f_X f_Y$, where f_X denotes the frequency of mononucleotide X in the sequence and f_{XY} the frequency of dinucleotide XY, etc. As a conservative criteron, for $\rho_{XY} > 1.25$ (or < 0.78), the XY pair is regarded to be of high (or low) relative abundance compared with a random association of mononucleotides (25). In the case of double-stranded DNA (dsDNA), we take the symmetrized frequency of mononucleotides as $f^*_A = f^*_T = (f_A + f_T)/2$, $f^*_C = f^*_G = (f_C + f_G)/2$, and $f^*_{GT} = f^*_{AC} = (f_{GT} + f_{AC})/2$ for the symmetrized double-stranded frequency of GT/AC, and so on. The symmetrized dinucleotide odds ratio measure is taken to be $\rho^*_{GT} = f^*_{GT}/f^*_G f^*_T = 2(f_{GT} + f_{AC})/(f_G + f_C)(f_T + f_A)$ and similarly for all dinucleotides (for rationale and justifications, see reference 8). For single-stranded viruses, we evaluate both single-stranded (ρ) and symmetrized double-stranded (ρ^*) relative abundance measures.

RESULTS

CpG deficiencies in small (<30-kb) viral genomes of vertebrate species. All completely sequenced eukaryotic viral genomes (Table 1) were examined for extremes of dinucleotide relative abundances (see Methods). With only a few exceptions (all in the togavirus family; see Discussion), all viral genomes are decisively CpG suppressed. This CpG deficit prevails independent of viral genomic organization and morphology (i.e., whether dsDNA, dsRNA, single-stranded [ssDNA], or ssRNA [of positive or negative polarity], enveloped or not). The ρ^*_{CG} measure of relative abundance does not correlate with size of genome or percent C+G content. Early observations of CpG suppression in a few small viruses were demonstrated experimentally by Subak-Sharp et al. (49), using the techniques of Josse et al. (30) (see also reference 39). Recently, Shpaer and Mullins (46) noted CpG suppression in lentiviruses.

Apart from CpG deficiencies, there are no other pervasive significant dinucleotide extremes. The relative abundance of GpC conforms with random expectations, i.e., the ρ^*_{GC} values hover about 1.00. The relative abundance of TpA (ρ^*_{TA}) among small viruses is (with one exception) less than 1, largely in the range 0.70 to 0.90. This conforms with the tendency of TpA underabundance for most prokaryotic and eukaryotic species (8).

CpG relative abundances in the different codon sites. The

* Corresponding author. Mailing address: Department of Mathematics, Building 380, Stanford University, Stanford, CA 94305. Phone: (415) 723-2204. Fax: (415) 725-2040. Electronic mail: fd.zgg@forsythe.stanford.edu.

TABLE 1. Dinucleotide relative abundances in short (<30-kb) eukaryotic viral genomes

Virus (type)[a]	Host	Genome length (nt)[b]	Base composition (%)					TpG/CpA[c]			CpC/GpG[c]			GpC[d]	CpG[d]
			A	C	G	T	C+G	ρ*	ρ	ρ	ρ*	ρ	ρ	ρ*	ρ*
Papovaviruses (circular dsDNA)															
Papillomavirus type 11	Human	7,931	30	19	22	29	41	1.35			1.21			1.03	0.46
Papillomavirus type 1	Bovine	7,945	29	22	24	26	46	1.24			1.07			1.18	0.46
BK virus Dun	Human	5,153	30	20	20	30	40	1.14			1.36			1.17	0.06
Papovavirus	Monkey	5,089	30	20	21	30	41	1.12			1.38			1.05	0.14
Papovavirus	Hamster	5,306	30	21	20	28	41	1.22			1.15			1.09	0.16
Simian virus 40	Monkey	5,243	29	21	20	30	41	1.29			1.20			1.22	0.12
JC polyomavirus	Human	5,130	30	20	20	30	40	1.24			1.29			1.22	0.08
Polyomavirus A2	Mouse	5,297	26	24	23	26	47	1.19			1.21			0.98	0.32
Murine polyomavirus	Mouse	4,754	31	20	21	29	41	1.15			1.22			1.19	0.34
Polyomavirus	Bovine	4,697	29	21	20	30	41	1.16			1.29			1.07	0.20
Hepadnaviruses (circular enveloped dsDNA)															
Hepatitis B virus	Human	3,188	22	27	22	28	49	1.11			1.24			0.85	0.55
Hepatitis B virus	Duck	3,027	29	23	21	27	44	1.05			1.19			0.98	0.58
Hepatitis B virus	Heron	3,027	29	24	21	26	45	1.05			1.23			0.87	0.59
Hepatitis B virus	Chimpanzee	3,182	24	27	21	28	48	1.14			1.22			0.91	0.54
Hepatitis B virus	Squirrel	3,311	27	23	20	30	43	1.17			1.24			0.94	0.44
Hepatitis B virus	Woodchuck	3,308	25	25	20	30	45	1.17			1.21			0.97	0.54
Parvoviruses (linear ssDNA, + or −)															
Adeno-associated virus	Human	4,675	26	27	27	21	54	1.14	1.11	1.17	1.03	0.98	1.08	0.90	0.80
Parvovirus CPV-N	Canine	5,323	36	16	20	28	36	1.22	1.26	1.22	1.20	1.07	1.26	0.96	0.44
Parvovirus ADV-G	Mink	4,801	36	19	20	25	39	1.20	1.33	1.11	1.11	0.98	1.22	0.90	0.36
Minute virus	Mouse	5,149	22	20	22	25	42	1.30	1.39	1.24	1.08	0.95	1.19	1.12	0.38
Retroviruses (linear enveloped ssRNA+)															
Lentiviruses															
HIV type 1	Human	9,229	36	18	24	22	42	1.17	1.10	1.30	1.24	1.28	1.17	1.05	0.20
HIV type 2	Human	11,443	33	21	25	21	46	1.19	1.16	1.25	1.22	1.23	1.20	1.02	0.26
SIV	Monkey[e]	9,597	33	21	24	22	45	1.17	1.09	1.27	1.23	1.20	1.25	1.00	0.35
SIV	Mandrill	9,215	36	16	24	24	40	1.13	1.07	1.27	1.20	1.24	1.11	1.00	0.14
FIV	Feline	9,474	38	14	22	25	36	1.08	1.09	1.15	1.30	1.42	1.17	0.99	0.28
Infectious anemia virus	Equine	8,407	35	16	22	26	39	1.15	1.15	1.20	1.30	1.22	1.29	0.90	0.28
Visna virus	Sheep	9,225	37	15	26	21	41	1.10	1.15	1.21	1.33	1.45	1.19	0.97	0.37
Other retroviruses															
Foamy virus	Monkey	12,972	32	19	20	28	39	1.11	1.21	1.02	1.23	1.31	1.17	1.00	0.32
HTLV I	Human	9,067	23	35	19	23	54	1.01	1.04	1.00	1.29	1.14	1.33	0.91	0.55
HTLV II	Human	8,952	24	36	18	22	54	1.01	1.03	0.98	1.37	1.14	1.60	0.75	0.50
SRV-1 type D	Monkey	8,173	31	24	19	27	43	1.08	1.17	1.00	1.31	1.26	1.34	1.02	0.46
Mason-Pfizer D monkey virus	Monkey	8,557	30	24	19	27	43	1.10	1.20	1.01	1.31	1.25	1.34	0.98	0.45
Jaagsiekte retrovirus	Sheep	7,462	28	22	20	30	42	1.07	1.11	1.03	1.26	1.22	1.29	1.07	0.64
Rous sarcoma virus	Avian	9,625	24	25	29	22	54	1.09	1.12	1.07	1.16	1.16	1.15	0.97	0.68
MMTV	Mouse	10,125	30	21	23	26	44	1.07	1.09	1.07	1.29	1.31	1.27	0.91	0.44
Friend leukemia virus	Mouse	8,323	25	29	24	21	53	1.03	1.13	0.94	1.28	1.27	1.27	0.78	0.51
Endogenous type C retrovirus	Human	8,342	29	22	24	25	46	1.14	1.24	1.06	1.29	1.42	1.19	0.92	0.28
Endogenous type C retrovirus	Baboon	8,018	27	29	23	22	51	0.98	1.05	0.92	1.31	1.25	1.36	0.82	0.54
Togaviruses (linear enveloped ssRNA+)															
Encephalitis virus	Equine	11,444	28	25	25	22	50	1.22	1.28	1.17	0.99	1.04	0.94	1.07	0.76
Sindbis virus	Human	11,703	28	26	25	21	51	1.16	1.20	1.12	0.95	1.02	0.88	1.06	0.90
O'nyong-nyong virus	Human	11,835	31	24	24	21	48	1.22	1.26	1.20	0.93	0.92	0.94	1.04	0.76
Ross River virus	Human	11,657	28	25	26	21	51	1.20	1.27	1.15	0.99	1.06	0.93	1.00	0.82
Semliki Forest virus	Human	11,442	27	26	27	20	53	1.18	1.26	1.13	0.95	0.97	0.93	1.01	0.89
Rubella virus	Human	9,755	15	39	31	15	70	1.17	1.29	1.08	0.90	0.91	0.85	1.12	1.04
Arteritis virus	Equine	12,687	21	26	26	27	52	1.31	1.32	1.30	0.98	0.99	0.96	1.13	0.75
Lelystad virus	Human	15,101	21	27	25	26	53	1.24	1.27	1.23	1.09	1.09	1.09	0.96	0.70
Lactate dehydrogenase virus	Human	14,225	23	23	26	28	49	1.34	1.38	1.26	1.10	1.23	1.00	1.01	0.59
Cholera virus	Hog	12,284	31	21	26	22	47	1.15	1.22	1.13	1.19	1.23	1.14	0.87	0.47
Picornaviruses (linear ssRNA+)															
Poliovirus type 2	Human	7,440	29	24	23	24	47	1.33	1.33	1.32	1.11	1.10	1.12	0.94	0.49
Coxsackievirus A24	Human	7,461	30	23	23	25	46	1.31	1.31	1.32	1.14	1.16	1.13	0.86	0.48
Enterovirus	Bovine	7,414	26	25	24	24	49	1.25	1.28	1.23	1.10	1.16	1.05	1.01	0.59
Vesicular stomatitis virus	Swine	7,400	28	24	25	22	49	1.33	1.33	1.34	1.06	1.04	1.08	0.92	0.62

Continued on facing page

TABLE 1—*Continued*

Virus (type)[a]	Host	Genome length (nt)[b]	Base composition (%)					TpG/CpA[c]			CpC/GpG[c]			GpC[d]	CpG[d]
			A	C	G	T	C+G	ρ*	ρ	ρ	ρ*	ρ	ρ	ρ*	ρ*
Hepatitis A virus	Human	7,478	29	16	22	33	38	1.31	1.25	1.39	1.23	1.27	1.16	0.86	0.14
Hepatitis A virus	Simian	7,400	29	16	22	33	38	1.32	1.24	1.41	1.18	1.16	1.15	0.85	0.15
Encephalomyocarditis virus	Human	7,825	26	25	24	25	49	1.19	1.32	1.08	1.20	1.34	1.03	0.86	0.56
Encephalomyelitis virus	Mouse	8,105	24	27	22	27	49	1.19	1.35	1.06	1.12	1.12	1.09	0.92	0.65
Rhinovirus type 1B	Human	7,133	33	18	19	29	37	1.33	1.35	1.31	1.16	1.18	1.14	0.94	0.25
Echovirus 22	Human	7,339	32	19	20	29	39	1.35	1.38	1.32	1.21	1.22	1.20	0.90	0.20
Flaviviruses (linear enveloped ssRNA+)															
Yellow fever virus	Human	10,862	27	21	28	23	49	1.37	1.47	1.30	1.14	1.19	1.09	0.86	0.38
Cell-fusing virus	Human	10,695	24	24	27	24	51	1.20	1.23	1.15	1.07	1.04	1.08	0.88	0.79
Kunjin virus	Human	10,664	27	22	29	22	51	1.32	1.42	1.25	1.06	1.08	1.02	0.92	0.53
West Nile virus	Human	10,960	27	23	28	22	51	1.32	1.44	1.24	1.02	1.04	0.99	0.91	0.57
Encephalitis virus	Human	10,976	28	23	28	21	51	1.27	1.36	1.22	1.05	1.05	1.03	0.95	0.58
Tick-borne encephalitis virus	Human	10,546	25	22	32	21	54	1.36	1.45	1.31	1.07	1.03	1.03	0.87	0.53
Dengue virus type 3	Human	10,696	32	21	26	21	47	1.31	1.37	1.32	1.13	1.11	1.12	0.90	0.41
Hepatitis C virus	Human	9,400	20	30	28	22	58	1.20	1.22	1.18	1.09	1.02	1.16	0.94	0.73
Viral diarrhea virus	Bovine	12,573	32	20	26	22	46	1.18	1.25	1.17	1.23	1.24	1.20	0.87	0.40
Caliciviruses (linear ssRNA+)															
Hemmorhagic fever virus	Rabbit	7,437	26	25	25	24	50	1.42	1.43	1.41	1.05	1.05	1.04	0.93	0.59
Calicivirus	Feline	7,690	27	23	23	27	46	1.28	1.28	1.18	1.10	1.05	1.15	0.97	0.60
Hepatitis E virus	Human	7,207	17	32	26	25	58	1.11	1.23	1.02	1.10	1.13	1.03	1.05	0.80
Norwalk agent	Human	7,644	28	23	25	24	48	1.29	1.34	1.25	1.22	1.25	1.20	0.92	0.47
Coronavirus (ssRNA+)															
Infectious bronchitis virus	Avian	27,608	29	16	22	33	38	1.26	1.25	1.24	0.96	1.00	0.91	1.16	0.49
Paramyxoviruses (linear enveloped ssRNA−)															
Measles virus	Human	15,894	29	24	23	23	47	1.19	1.17	1.21	1.19	1.20	1.18	0.85	0.49
Parainfluenza virus	Human	10,603	25	24	21	30	45	1.17	1.19	1.15	1.07	1.11	1.00	0.84	0.49
Orthomyxoviruses (linear enveloped ssRNA−)															
Influenza A virus	Human	13,606	33	19	24	23	43	1.28	1.34	1.28	1.07	1.05	1.06	0.93	0.44
Influenza B virus	Human	14,613	36	18	22	24	40	1.26	1.40	1.19	1.20	1.19	1.18	1.02	0.34
Influenza C virus	Chicken	14,071	36	17	21	26	38	1.22	1.33	1.17	1.12	1.18	1.06	1.10	0.31
Rhabdoviruses (linear enveloped ssRNA−)															
Rabies virus	Vertebrates	11,928	29	22	23	26	45	1.12	1.12	1.11	1.15	1.15	1.14	0.67	0.46
Vesicular stomatitis virus	Bovine	11,161	31	20	22	27	42	1.21	1.21	1.21	1.15	1.14	1.15	0.77	0.48
Large viruses															
Adenoviruses (linear dsDNA)															
Adenovirus type 2	Human	35,937	23	28	27	22	55	1.12			1.01			1.16	0.89
Adenovirus type 12	Human	34,125	27	23	23	26	46	1.12			1.06			1.20	0.81
Poxvirus (linear enveloped dsDNA)															
Vaccinia virus	Human	191,737	33	17	17	33	34	0.99			1.00			0.80	1.11
Herpesviruses (linear enveloped dsDNA)															
HSV type 1	Human	152,260	16	34	34	16	68	0.98			1.07			0.93	1.01
Varicella-zoster virus	Human	124,884	27	23	23	27	46	1.00			1.18			0.94	1.14
Cytomegalovirus	Human	229,354	22	28	29	21	57	1.05			0.86			1.08	1.19
EBV	Human	172,281	20	30	29	20	59	1.13			1.21			0.90	0.60
Herpesvirus 1	Equine	150,223	22	29	28	22	57	1.01			1.02			1.05	0.99
Herpesvirus saimiri	Monkey	112,930	33	18	16	32	34	1.23			0.87			1.28	0.33
Herpesvirus 1	Catfish	134,226	21	28	28	22	56	0.98			1.04			0.71	1.11

[a] HIV, human immunodeficiency virus; SIV, simian immunodeficiency virus; FIV, feline immunodeficiency virus; HTLV, human T-cell leukemia virus; SRV, simian retrovirus; MMTV, mouse mammary tumor virus; HSV, herpes simplex virus.

[b] nt, nucleotides.

[c] Single-strand relative abundance values (ρ) are listed first for the upper dinucleotide and then for the lower dinucleotide.

[d] Single-strand ρ values are not shown because of a very high concordance with the symmetrized ρ* values, typically differing by ≤1%.

[e] Host is African green monkey.

TABLE 2. Relative abundances (ρ) of CpG dinucleotides in retrovirus *gag*, *pol*, and *env* genes

Host	Virus[a]	ρ for:				
		Codon positions			Total coding	Complete genome
		1 and 2	2 and 3	3 and 1		
Human	HIV type 1	0.20	0.18	0.15	0.18	0.20
Human	HIV type 2	0.32	0.33	0.17	0.27	0.26
Monkey	SIV	0.30	0.34	0.27	0.30	0.35
Mandrill	SIV	0.11	0.10	0.10	0.10	0.14
Feline	FIV	0.23	0.20	0.16	0.20	0.30
Equine	Infectious anemia virus	0.21	0.25	0.27	0.24	0.28
Sheep	Visna virus	0.45	0.44	0.25	0.38	0.39
Monkey	Foamy virus	0.31	0.26	0.32	0.29	0.32
Human	HTLV I	0.47	0.52	0.57	0.53	0.60
Human	HTLV III	0.25	0.14	0.13	0.17	0.21
Monkey	SRV-1 type D	0.36	0.43	0.37	0.39	0.47
Monkey	Mason-Pfizer D	0.32	0.41	0.37	0.37	0.46
Sheep	Jaagsiekte retrovirus	0.74	0.61	0.71	0.67	0.64
Avian	Rous sarcoma virus	0.73	0.56	0.74	0.68	0.69
Mouse	MMTV	0.41	0.39	0.44	0.41	0.44
Mouse	Leukemia virus	0.56	0.47	0.45	0.49	0.51
Baboon	Endogenous type C	0.54	0.45	0.55	0.51	0.54

[a] See Table 1, footnote *a*, for abbreviations.

TABLE 3. Arginine codon usages[a]

Viral family	No. of codons		
	CGN	AGR	Total
Retrovirus	795	1,915	42,521
Hepadnavirus	204	211	6,921
Parvovirus	155	204	9,170
Papovavirus	79	400	13,337
Togavirus[b]	442	289	10,917
Picornavirus	153	391	10,963
Flavivirus	213	595	14,224
Calicivirus	48	62	2,462
Paramyxovirus	130	278	5,463
Orthomyxovirus	159	702	13,534
Rhabdovirus	142	326	7,144

[a] Counts of CGN and AGR are aggregates of all viruses within each family for which coding sequence annotations are included in GenBank 77.

[b] Togavirus counts are dominated by rubella virus, containing 297 CGN and 48 AGR codons.

aggregate dinucleotide frequencies at codon sites 1 and 2, 2 and 3, and 3 and 1 of the *gag*, *pol*, and *env* genes in each retrovirus were evaluated with respect to CpG relative abundances (Table 2). Independent of the codon site pairings, we observed approximately uniform CpG suppression. This argues against directed mutation and/or selection bias with respect to amino acid usages and with respect to synonymous versus nonsynonymous substitutions in accounting for CpG deficits.

Arginine usage in relation to CpG suppression. Is there a significantly low level of usage of arginine in small viral proteins compared with that in host proteins? To what extent are arginine residues encoded from CGN codons vis-à-vis AGR codons? The answer to the first question is no. For example, arginine usage in human proteins is on average 5.3% (31), whereas average arginine usage is 5.6% in aggregate papovaviruses, 6.7% in hepadnaviruses, 4.2% in parvoviruses, 5.7% in flaviviruses, and 5.0% in paramyxoviruses. However, in encoding arginine, AGR (two codon complements) is generally used by a factor of 2 more than CGN (four codon complements) in small viral genomes (Table 3). In contrast, there is no bias in codon preferences for arginine in large viruses. In fact, in the Epstein-Barr virus (EBV), the average frequency of arginine encoded by CGN codons is 8.0% and that by AGR is 5.4%; in varicella-zoster virus the average frequency of CGN is 4.8% and that of AGR is 1.6%; in herpes simplex virus type 1 the corresponding frequencies are 7.8 and 0.7%, respectively; and the frequencies are similar for the other large viruses (data not shown). Even with adjustment for large viral genome composition, there is no bias in arginine codon usage of CGN versus AGR.

Small plant viruses also tend to be CpG suppressed. With only two exceptions (foxtail mosaic virus [$\rho^*_{CG} = 0.90$] and shallot virus X[$\rho^*_{CG} = 0.82$]), ρ^*_{CG} values for all plant viruses are significantly low (Table 4).

No consistent CpG suppression is observed in large viral genomes of vertebrate species. Examination of Table 1 shows that except for the gammaherpesvirus class (EBV and herpesvirus saimiri), there is no significant over- or underrepresentation of CpG (see also reference 32). Although the moderate-size adenovirus type 2 and type 12 genomes do not show CpG relative abundance as significantly low, these viruses tend to the low side in CpG representations (Table 1). In contrast, the larger herpesviruses (excluding the gamma types) and vaccinia virus carry normal to moderately high ρ^*_{CG} relative abundance values. The highest ρ^*_{CG} value among the viruses occurs for the human cytomegalovirus ($\rho^*_{CT} = 1.19$), which has a broad cellular range in its latent state (38).

For comparison, all prokaryotic phages examined carry CpG dinucleotides in the normal relative abundance range. Interestingly, the temperate phages (e.g., λ, Mu, P1, P4, and P22) and several others exhibit significantly high relative abundance of the reverse dinucleotide GpC, and this is also valid for many bacterial genomes (6).

Dinucleotide relative abundance extremes for retrotransposable elements. To some extent, retrotransposons can be considered endogenous retrovirus-like elements involving terminal direct repeats, often containing a reverse transcriptase-like gene. Retrotransposons are mobile through an RNA intermediate. All eukaryotic retrotransposons (so designated in GenBank version 77) exceeding 5 kb were evaluated for compositional extremes. The retrovirus genomes of Table 1 invariably exceed 7 kb in length, the largest reaching 12 kb. Additionally, all available elements in the *Drosophila melanogaster* data bank characterized as any kind of transposon were examined with respect to CpG representations. The ρ^*_{CG} values are reported in Tables 5 and 6. The *copia* sequence is significantly CpG suppressed, and several other *copia*-like elements reveal low ρ^*_{CG} values. There have been reports of an encapsulated *copia* resembling a bona fide retrovirus (19). Most of the retrotransposons in Table 5 have low CpG relative abundances. However, smaller transposable vestiges (≤5 kb) are less predictable with respect to CpG relative abundances (data not shown).

DISCUSSION

We first present some issues and then venture some interpretations and speculations. How do methylation and other DNA modifications affect CpG representations in viral genomes? Except for tRNAs, essentially nothing is known about methylation of RNA, including RNA viruses. Can events of viral integration and excision to and from the host genome for DNA viruses influence the level and distribution of CpG

TABLE 4. Dinucleotide relative abundances in plant virus genomes

Virus and genome type	Genome length (nt)[a]	Base composition (%)					TpG/CpA			CpC/GpG			GpC	CpG
		A	C	G	T	C+G	ρ^*	ρ	ρ	ρ^*	ρ	ρ	ρ^*	ρ^*
dsDNA														
Commelina yellow mottle	7,516	35	19	21	25	40	1.19			1.02			1.02	0.44
Soybean chlorotic mottle	8,203	41	17	17	25	34	1.02			1.10			0.97	0.55
Figwort mosaic	7,743	39	17	18	26	35	1.03			1.09			0.99	0.53
Southern bean mosaic	4,194	23	25	26	25	52	1.12			1.08			0.94	0.74
ssDNA														
Beet curly top	2,994	29	17	23	32	39	1.07			1.17			0.71	0.65
ssRNA														
Tobacco necrosis	2,759	27	21	25	26	46	1.20	1.17	1.25	1.15	1.25	1.06	0.91	0.60
Parsnip yellow fleck	9,871	29	20	23	28	43	1.22	1.18	1.26	1.10	1.11	1.09	1.04	0.43
Tomato bushy stunt	4,803	26	21	27	25	48	1.15	1.13	1.18	1.05	0.95	1.08	0.85	0.71
Artichoke mottled crinkle	4,816	26	20	28	26	48	1.13	1.13	1.14	1.07	1.03	1.06	0.87	0.71
Cucumber necrosis	4,701	26	21	27	25	49	1.14	1.15	1.13	1.10	1.03	1.11	0.92	0.73
Cymbidium ringspot	4,733	26	22	28	24	50	1.14	1.11	1.18	1.11	1.01	1.15	0.83	0.71
Turnip yellow mosaic	6,318	23	39	17	21	56	1.09	1.00	1.09	1.13	0.94	1.17	0.69	0.78
Eggplant mosaic	6,331	21	39	16	25	54	1.15	1.30	1.14	1.14	0.93	1.19	0.78	0.54
Kennedya mosaic	6,362	23	39	15	23	54	1.07	1.11	1.07	1.23	0.99	1.27	0.65	0.57
Ononis yellow mosaic	6,237	21	35	16	28	51	1.10	1.21	1.13	1.17	0.97	1.26	0.74	0.56
Apple stem grooving	6,496	31	18	23	28	41	1.19	1.18	1.21	1.12	1.11	1.10	0.91	0.49
Potato M	8,535	26	20	28	25	49	1.28	1.32	1.24	0.90	0.89	0.87	1.25	0.73
Shallot X	8,832	29	29	21	22	49	1.18	1.22	1.11	0.95	0.90	0.98	1.02	0.82
Apple chlorotic leaf spot	7,555	32	18	24	27	42	1.26	1.19	1.38	1.10	1.02	1.11	0.88	0.41
Potato X	6,435	31	24	23	23	47	1.28	1.26	1.29	1.00	0.97	1.03	0.99	0.49
Foxtail mosaic	6,151	27	30	23	20	52	1.16	1.19	1.11	0.94	0.87	1.03	0.99	0.90
Papaya mosaic	6,656	30	25	23	22	48	1.17	1.16	1.16	1.20	1.19	1.22	0.90	0.50
White clover mosaic	5,846	30	27	17	26	44	1.25	1.45	1.11	1.08	1.02	1.04	0.90	0.50
Strawberry mild yellow	5,966	25	28	23	24	51	1.14	1.26	1.05	1.03	1.06	0.97	0.96	0.74
Potato Y	9,704	31	19	23	27	42	1.33	1.30	1.39	0.92	0.87	0.93	1.11	0.58
Papaya ringspot	10,352	31	18	24	27	42	1.25	1.26	1.26	0.91	0.86	0.92	1.05	0.73
Pea seed-borne mosaic	9,924	33	18	24	27	42	1.37	1.29	1.51	0.89	0.75	0.93	1.11	0.63
Plum pox	9,741	31	20	23	25	43	1.35	1.30	1.41	0.88	0.77	0.96	1.03	0.70
Tobacco etch	9,497	31	19	24	25	43	1.35	1.32	1.42	0.93	0.82	0.97	1.14	0.58
Tobacco vein mottling	9,472	32	19	23	26	42	1.33	1.26	1.42	0.89	0.79	0.95	1.22	0.62
Pepper mottle	9,666	32	18	23	27	41	1.32	1.30	1.37	0.93	0.84	0.96	1.08	0.65
Barley yellow	5,677	29	24	25	22	48	1.15	1.18	1.13	1.04	1.05	1.03	0.95	0.77
Tobacco mosaic	6,355	30	17	24	29	41	1.15	1.19	1.10	0.98	1.00	0.95	0.97	0.71

[a] nt, nucleotides.

occurrences? Are there structural and/or regulatory constraints intrinsic to the CpG dinucleotide? These might be different for DNA and RNA strands. Can a given CpG content or pattern of distribution cause a certain structure, particularly in the interaction with specific proteins? Are there nonmethylation mechanisms, preferential mutation, or other selective forces specific to CpG dinucleotides? Is there an evolutionary founder component to CpG suppression?

Methylation. Prokaryotic methylation by a restriction system directed at a specific restriction site can prevent cleavage by the corresponding endonuclease, or a different methylation process can be (and is) used by DNA repair systems to distinguish between old and new DNA chains. By contrast, the purpose and function of CpG methylation in eukaryotes are not well understood. It is thought that cytosine methylation is essential for mammalian development, epigenetic gene regulation, and the maintenance of X-chromosome inactivation in mammals (2, 4, 21, 26, 44, 52).

Methylation can influence promoter activity and might also affect DNA replication, recombination, repair, and transposition (13, 15, 16). The 5-methylcytosine positions apparently participate as hot spots of recombination and mutagenesis. For example, methylation at cytosine bases seems to increase the mutation rate by at least 1 order of magnitude in some human genes, as witnessed in Rb1 and p53, which may stimulate carcinogenesis (11, 12, 18, 29, 42). On the other hand, it has been proposed elsewhere (15) that methylation can function in part as a mechanism of defense against uptake, integration, and expression of foreign DNA.

CpG methylation is often correlated with reduction of gene activity. Along these lines, methylation of promoter sequences can cause their assembly into condensed inactive chromatin and strongly inhibits transcription in vitro and in vivo (for example, see references 2, 13, 15, and 44). In particular, methylation keeps chromatin in a condensed state, whereas unmethylated sequences are more relaxed and accessible to diffusible factors (33). Protection against methylation can be effected by protein-specific binding to appropriate sites (44). This is reminiscent of nucleosome placements prevented by competition from protein binding to relevant sites (47).

Small-virus CpG suppression generally may not be a result of methylation-mutation. Examination of Table 1 reveals that in several cases of DNA viruses or viruses having a DNA intermediate (e.g., with retroviruses and hepadnaviruses [dsDNA]), the relative abundance of TpG/CpA is only slightly above average, in the normal range ($1.00 \leq \rho^*_{TG/CA} \leq 1.20$),

TABLE 5. Dinucleotide relative abundances in selected (≥5-kb) retrotransposons

Host	Retrotransposon	Length (nt)[a]	Base composition (%)					ρ* representation			
			A	C	G	T	C+G	TpG/CpA	CpC/GpG	GpC	CpG
Human	LINE-1	6,065	39	21	20	19	41	1.23	1.22	0.96	0.32
Mouse	Virus-like BVL-1	5,447	26	22	24	28	46	1.12	1.13	0.78	0.43
Silkworm	Insertion R2	6,558	22	25	31	21	56	1.06	1.01	0.93	0.95
D. silvestris	U28T2 21	6,304	30	23	23	23	46	1.07	1.09	1.06	0.84
	U28T2 23	5,243	28	25	27	20	52	1.12	1.12	1.03	0.66
	U28T2 24	7,779	29	23	24	23	47	1.09	1.11	1.04	0.78
D. virilis	Ulysses	10,653	27	22	27	23	49	1.01	1.03	1.11	1.07
S. cerevisiae	Ty3	5,530	35	24	17	24	41	1.07	1.05	0.78	0.93
	Ty4	7,654	39	16	17	28	34	1.16	1.04	0.73	0.75
	Ty4A	7,000	39	16	18	27	42	1.18	1.00	0.79	0.67
D. discoideum	DRE	6,428	49	18	11	21	39	1.22	1.16	0.77	0.65
	DIRS-1	7,053	34	20	16	29	36	1.14	0.99	0.74	0.82
A. thaliana	*copia*-like Tal-3	5,258	32	16	25	27	41	1.22	0.88	0.83	0.46
C. fulvum	Cft-1	7,396	31	26	24	19	50	1.03	0.90	0.86	0.98
Lilium henryi	del	9,345	31	19	17	33	36	1.02	1.38	0.77	0.40
Potato	*copia*-like	5,060	32	17	21	30	38	1.13	1.15	0.80	0.40

[a] nt, nucleotides.

contrary to expectations under the methylation-deamination-mutation scenario. In contrast, other single-stranded positive-strand RNA viruses (picornaviruses, flaviviruses, caliciviruses, and coronaviruses) which are probably not methylated carry significantly high TpG/CpA relative abundances. Apropos, simian virus 40 as a free particle was observed to be in an unmethylated state (13, 16). An exception is frog virus 3, an iridovirus: its genome is >20% methylated and mainly in CpG (43, 55). Lentivirus genomes are not methylated prior to integration into the host DNA (46). Similarly, free adenovirus type 12 is unmethylated and only after integration into the host genome does de novo methylation initiate (14). It has also been established that the retrovirus murine leukemia virus is unmethylated during conversion (reverse transcription) to dsDNA in the cytoplasm and becomes methylated only 8 to 16 days after infection (22, 40). Among the retroviruses, the lentiviruses are more strongly CpG suppressed than other retroviruses, with $0.14 \leq \rho^*_{CG} \leq 0.37$ (average, about 0.26) in the former group and $0.28 \leq \rho^*_{CG} \leq 0.68$ (average, about 0.49) in the latter group (Table 1). The mechanism is unlikely to be methylation since the relative abundances of TpG/CpA are quite normal. Viral latency can be associated with methylation of the viral genome (56). Why a free dsDNA viral genome concentrated in the nucleus should not be methylated is unclear. Does methylation require a special DNA-protein or chromatin structure? In some circumstances, it is conceivable that small viruses can seclude themselves, avoiding methylation.

The gammaherpesviruses (e.g., EBV, herpesvirus saimiri, and bovine herpesvirus 4) are potently CpG suppressed and tend to have high TpG/CpA relative abundances. It is unknown whether CpG dinucleotides are methylated during replication of gammaherpesvirus sequences. Various degrees of methylation in different tumorigenic cell lines of EBV have been detected, ranging from unmethylated to an extensively methylated state, with concomitant differential EBV latent gene expression depending on the degree of methylation (28, 35, 37). In its most stable latent state, EBV tends to be unmethylated (28). Honess et al. (27) propose that the standard methylation-deamination-mutation hypothesis applies to herpesviruses found in highly dividing cells such as B and T lymphocytes. However, Marek's disease virus and human herpesvirus 6, which present some biological features of gammaherpesviruses (e.g., they are primarily lymphotropic), do not show CpG suppression or any biased dinucleotide relative

TABLE 6. Dinucleotide relative abundances in selected *D. melanogaster* transposons

Gene	Length (nt)[a]	Base composition (%)					ρ* representation			
		A	C	G	T	C+G	TpG/CpA	CpC/GpG	GpC	CpG
copia-like retroposons										
copia	5,183	36	14	19	31	33	1.17	1.04	1.20	0.69
297	6,995	40	19	14	27	33	1.10	1.21	0.97	0.83
gypsy	2,790	31	25	21	23	46	1.09	1.03	0.99	0.82
17.6	7,439	40	20	14	25	34	1.24	1.10	1.08	0.71
412	6,897	40	18	17	25	35	1.11	1.06	1.23	0.80
vlp	4,960	36	14	19	30	33	1.17	1.03	1.20	0.70
Other transposons (with short terminal repeats)										
P	2,889	35	17	19	30	36	1.07	1.04	1.21	0.98
hobo	3,016	34	19	19	28	38	1.09	1.11	1.13	0.88
pogo	2,121	34	17	19	30	36	1.22	0.79	1.73	1.01

[a] nt, nucleotides.

abundances (32). DNA methylation may be feasible only in the presence of a particular DNA or chromatin context.

Mechanisms leading to CpG deficits in small viruses might be different from mechanisms effecting CpG suppression via standard methylation. The retroviruses are uniformly CpG suppressed independent of codon sites (Table 2). Apropos, all metazoan species (vertebrate and invertebrate) mitochondrial genomes are without exception strongly CpG suppressed, however, with relative frequencies of TpG/CpA mostly in the normal range (9). Most retrotransposons are also CpG suppressed. In the mitochondrial context, it is unlikely that methylation is involved since invertebrates (e.g., *Drosophila melanogaster*, *C. elegans*, and sea urchin) apparently do not possess the standard methyltransferase, and in vertebrates the methyltransferase cannot access the mitochondrion. In this light, there appear to be other mechanisms or factors which cause CpG depletion, at least in dsDNA sequences. Is it possible that these dinucleotides are part of an important regulatory or structural sequence whose frequency should be kept distinctly low for optimum functioning? An example pertains to the dinucleotide TpA. It is established that TpA is intrinsically less stable than all other dinucleotides (7, 12a). Evidence of substantial untwisting and bending at TpA steps occurs in transcription initiation via protein binding to the TATA box, with *Eco*RV bound to its recognition sequence (GATATC) and resolvase bound at the site at which crossing-over occurs (e.g., see reference 54). A general thesis suggests that protein-DNA complexes biochemically exploit the reduced thermodynamic stability of the TpA base pair. Whether CpG dinucleotides reflect on corresponding regulatory or structural capacities is unknown. It is demonstrated that CpG among dinucleotides possesses the highest thermodynamic stacking energy, at least 20% higher than that of GpC and CpC/GpG (7). CpG avoidance (with concomitant reduced dinucleotide stacking energy) putatively enhances the rate of transcription and replication of viral DNA, both activities requiring facile local strand separation and greater accessibility to host factors. From this perspective, low rates of CpG occurrence would be advantageous for a small genomic sequence. There is experimental evidence (3) that UpA is the RNA dinucleotide most susceptible to RNase activity. Could there be a corresponding defect in CpG doublets of RNA strands? It is also conceivable that ssDNA and ssRNA viruses in part have low CpG relative abundances in order to diminish secondary structure formations carrying CpG stackings that are difficult to disengage.

Drake et al. (17) argued that the spontaneous mutation rates among living organisms are inversely correlated with the sizes of the genomes. Many DNA viruses have high mutation frequencies and broad adaptability (48). Furthermore, RNA viruses are especially error prone, a fact generally attributed to the absence of RNA proofreading and mismatch-repair systems. Concomitantly, viral RNA genomes are ubiquitous cellular parasites that tend to replicate efficiently and can evolve extremely rapidly (24). Practically nothing is known about the methylation of RNA virus genomes that do not have a DNA intermediate, and therefore CpG suppression in these viruses cannot presently be readily related to methylation. RNA viruses are not expected to be under the same constraints as DNA viruses. Examination of Tables 1, 4, 5, and 6 reveals general tendencies of overrepresentations of TpG/CpA and CpC/GpG. It is tempting to speculate that for small mobile DNA or RNA sequences, including viruses, there exists a mechanism of preferential mutation from CpG to either TpG/CpA or CpC/GpG. Other processes of mutations include repeat induced point mutations, "RIPping" or RIP-like interactions which putatively help maintain a streamlined genome (34, 45). From another perspective, if CpG is a hot spot of mutation with potentially deleterious functional or structural consequences for proteins, a reduction in CpG occurrence would be selectively favorable for virus viability.

Runs of CpG (or, equivalently, runs of GpC) are known to be the ideal arrangement for inducing Z-DNA structures and/or left-hand superhelicity (41). It is, a priori, conceivable that CpG suppression serves to avoid these structural anomalies, although this seems unlikely because GpC shows generally normal representations.

Viral integration and excision from the host genome. It is familiar retrovirus biology for the virus to insert itself into the host genome and make RNA copies of itself. In this process, there are obvious advantages to low CpG relative frequencies that reduce methylation possibilities and concomitant transcriptional inhibitions of viral gene expression. Along these lines, methylation of several murine, feline, and avian retroviral proviruses correlates negatively with their expression (reviewed in reference 11). A similar succession putatively applies to the parvovirus family (1).

There are several well-studied examples in which a free DNA virus genome maintains an unmethylated state but upon integration into the host genome undergoes de novo methylation; e.g., adenovirus types 2 and 12 (15, 16, 50). Actually, de novo methylation and loss of methylation are both observed for cellular DNA abutting foreign viral DNA (15). Although incorporation of DNA virus genomes into host genomes is generally considered a dead end for the virus, their genomic sequences can remain intact through many replication cycles (for examples, see references 10 and 53). It is conceivable that on occasion an inserted sequence can be excised or produce recombinant progeny among incorporated sequences and/or with free viral sequences.

Coding versus noncoding CpG suppression. Comparing the ρ_{CG} relative abundances for coding regions of retroviruses with the ρ_{CG} values of the complete genome shows that there is almost always greater CpG suppression in the coding regions vis-à-vis the complete genome (Table 2). Small viral genomes tend to be streamlined. By contrast, the larger viruses seem to be capable of numerous DNA alterations, such as amplifications, excisions, inversions, and transpositions. In fact, many herpesviruses are strewn with substantial direct and inverted repeat sequences and also undergo lateral transfer of DNA between different hosts and other viruses (e.g., see references 5, 20, and 36).

There are four exceptions to CpG suppression among the small viruses, all restricted to togaviruses (Table 1). The rubella virus particularly stands out ($\rho^*_{CG} = 1.04$). The rubella virus is extraordinary in having a C+G content of $\approx$70%, whereas all other small viral genomes have C+G frequencies of $\leq$55%. In rubella virus proteins, the incidence of arginine encoded from CGN codons is dramatically high (Table 3). Maybe the need of arginine for rubella virus viability supersedes the possible difficulties or other purposes attendant to CpG occurrences.

Rubella virus, with a genome organization similar to that of the togavirus classification, is an outlier in many respects. For example, it is not arthropod borne (does not grow in insect cells) and infection is minimally cytopathic with persistence for many years. Sindbis and Semliki Forest viruses, the other exceptional non-CpG-suppressed viruses, are among the least virulent alphaviruses of the togavirus family (23). They both replicate primarily in mosquitoes rather than in human hosts. This life process and the type of mosquito may embody selective pressures with respect to genomic organization and

composition. Like rubella virus, the Sindbis and Semliki Forest viruses encode arginine predominantly from CGN versus AGR codons, in contrast to arginine codon preferences of other small viruses.

Is there an evolutionary founder component to CpG suppression? Parallels and contrasts between vertebrate methylation consequences and prokaryotic restriction systems and DNA repair methylase factors have been made (for examples, see references 13 and 44). We know that most phages are not CpG deficient so any historical factor is probably rooted in vertebrate evolution (6, 8). It is conceivable that early (500 to 600 million years ago) vertebrate viruses coevolved with vertebrates under conditions of reduced CpG genomic content. It is also conceivable that many of the vertebrate viruses have been derived from host DNA or RNA sources that retain the approximate CpG representations of the host. This cannot explain the preponderant low CpG representations in viruses of plant species. Most plant viruses are small RNA types (Table 4). The inverse correlation of degree of CpG suppression with genome size (in viruses, mitochondria [9], and mobile elements) presents an intriguing conundrum, and obvious experimental manipulations of the level and distribution of CpG dinucleotides might help elucidate their special role in these genomes.

ACKNOWLEDGMENTS

We appreciate the valuable discussions on the manuscript with C. Burge. We also gratefully acknowledge the helpful comments on the manuscript by P. Berg, B. Edwin Blaisdell, V. Brendel, A. M. Campbell, and E. S. Mocarski.

This work was supported in part by NIH grants HG00335-06 (S.K.), GM10452-29 (S.K.), and HG00085-01 (L.R.C.) and by NSF grant DMS91-06974. Walter Doerfler is indebted to Uta Francke and Paul Berg at Stanford for their hospitality during a sabbatical semester during the summer of 1993 and to the Volkswagen-Stiftung for financial support.

REFERENCES

1. **Berns, K. I., and M. A. Labow.** 1987. Parvovirus gene regulation. J. Gen. Virol. **68**:601–614.
2. **Bestor, T. H., and A. Coxon.** 1993. The pros and cons of DNA methylation. Curr. Biol. **6**:384–386.
3. **Beutler, E., T. Gelbart, J. Han, J. A. Koziol, and B. Beutler.** 1989. Evolution of the genome and the genetic code: selection of the dinucleotide level by methylation and polyribonucleotide cleavage. Proc. Natl. Acad. Sci. USA **86**:192–196.
4. **Bird, A. P.** 1993. Imprints on islands. Curr. Biol. **3**:275–277.
5. **Birkenbach, M., K. Josefsen, R. Yalamanchili, G. E. Lenoir, and E. Kieff.** 1993. Epstein-Barr virus-induced genes: first lymphocyte-specific G-protein-coupled peptide receptors. J. Virol. **67**:2209–2220.
6. **Blaisdell, B. E., A. M. Campbell, and S. Karlin.** Unpublished data.
7. **Breslauer, K. J., R. Frank, H. Blöcker, and L. A. Marky.** 1986. Predicting DNA duplex stability from the base sequence. Proc. Natl. Acad. Sci. USA **83**:192–196.
8. **Burge, C., A. M. Campbell, and S. Karlin.** 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. USA **89**:1358–1362.
9. **Cardon, L. R., C. Burge, D. Clayton, and S. Karlin.** Conundrum of CpG suppression in metazoan mitochondrial genomes. Proc. Natl. Acad. Sci. USA, in press.
10. **Challberg, M. D., and T. J. Kelly.** 1989. Animal virus DNA replication. Annu. Rev. Biochem. **58**:671–717.
11. **Cooper, D. N.** 1983. DNA methylation. Hum. Genet. **64**:315–333.
12. **Cooper, D. N., and H. Youssoufian.** 1988. The CpG dinucleotide and human genetic disease. Hum. Genet. **78**:151–155.

12a. **Delcourt, S. G., and R. D. Blake.** 1991. Stacking energies in DNA. J. Biol. Chem. **266**:15160–15169.

13. **Doerfler, W.** 1983. DNA methylation and gene activity. Annu. Rev. Biochem. **52**:93–124.
14. **Doerfler, W.** 1991. Abortive infection and malignant transformation by adenoviruses: integration of viral DNA and control of viral gene expression by specific patterns of DNA methylation. Adv. Virus Res. **39**:89–128.
15. **Doerfler, W.** 1991. Patterns of DNA methylation—evolutionary vestiges of foreign DNA inactivation as a host defense mechanism. Biol. Chem. Hoppe-Seyler **372**:557–564.
16. **Doerfler, W.** 1993. Patterns of de novo DNA methylation and promoter inhibition: studies on the adenovirus and the human genomes, p. 262–299. *In* J. P. Jost and H. P. Saluz (ed.), DNA methylation: molecular biology and biological significance. Birkhauser Verlag, Basel.
17. **Drake, J. W., E. F. Allen, S. A. Forsberg, R. M. Perparata, and E. O. Greening.** 1969. Spontaneous mutations. Nature (London) **221**:1128–1132.
18. **Fearon, E. R., and B. Vogelstein.** 1990. A genetic model for colorectal tumorigenesis. Cell **61**:759–767.
19. **Flavell, A. J., and C. Brierley.** 1986. The termini of extrachromosomal linear copia elements. Nucleic Acids Res. **14**:3659–3669.
20. **Fleckenstein, B., J. Albrecht, I. Müller-Fleckenstein, B. Biesinger, A. Ensser, and H. Fickenscher.** 1993. Rhadinoviruses and T-cell transformation, p. S-23. Abstr. Int. Herpesvirus Meeting, Pittsburgh, Pa., 23 to 30 July 1993.
21. **Gartler, S. M., and A. D. Riggs.** 1983. Mammalian X-chromosome inactivation. Annu. Rev. Genet. **17**:155–190.
22. **Gautsch, J. W., and M. C. Wilson.** 1983. Delayed de novo methylation in teratocarcinoma suggests additional tissue-specific mechanisms for controlling gene expression. Nature (London) **301**:32–37.
23. **Griffin, D. E.** 1986. Alphavirus pathogenesis and immunity, p. 209–249. *In* S. Schlesinger and M. J. Schlesinger (ed.), The Togaviridae and Flaviviridae. Plenum Press, New York.
24. **Holland, J.** 1993. Replication error, quasispecies populations and extreme evolutionary rates of RNA viruses, p. 203–218. *In* S. S. Morse (ed.), Emerging viruses. Oxford, New York.
25. **Hollander, M., and D. A. Wolfe.** 1973. Nonparametric statistical methods. Wiley, New York.
26. **Holliday, R.** 1987. The inheritance of epigenetic defects. Science **238**:163–170.
27. **Honess, R. W., U. A. Gompels, B. G. Barrell, M. Craxton, K. R. Cameron, R. Staden, Y. N. Chang, and G. S. Hayward.** 1989. Deviations from expected frequencies of CpG dinucleotides in herpesvirus DNAs may be diagnostic of differences in the states of their latent genomes. J. Gen. Virol. **70**:837–855.
28. **Jansson, A., M. Masucci, and L. Rymo.** 1992. Methylation of discrete sites within the enhancer region regulates the activity of the Epstein-Barr virus *Bam*HI W promoter in Burkitt lymphoma lines. J. Virol. **66**:62–69.
29. **Jones, P. A., W. M. Rideout III, J. C. Shen, C. M. Spruck, and Y. C. Tsai.** 1992. Methylation, mutation and cancer. Bioessays **14**:33–36.
30. **Josse, J., A. D. Kaiser, and A. Kornberg.** 1961. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J. Biol. Chem. **236**:864–875.
31. **Karlin, S., B. E. Blaisdell, and P. Bucher.** 1992. Quantile distributions of amino acid usage in protein classes. Protein. Eng. **5**:729–738.
32. **Karlin, S., E. S. Mocarski, and G. A. Schachtel.** 1994. Molecular evolution of herpesviruses: genomic and protein sequence comparisons. J. Virol. **68**:1886–1902.
33. **Keshet, I., J. Lieman-Hurwitz, and H. Cedar.** 1986. DNA methylation affects the formation of active chromatin. Cell **44**:535–543.
34. **Kricker, M. C., J. W. Drake, and M. Radman.** 1992. Duplication-targeted DNA methylation and mutagenesis in the evolution of eukaryotic chromosomes. Proc. Natl. Acad. Sci. USA **89**:1075–1079.
35. **Li-Fu, H., J. Minarovitz, C. Shi Long, B. Contreras-Salazar, L. Rymo, K. Falk, G. Klein, and I. Ernberg.** 1991. Variable expression of latent membrane protein in nasopharyngeal carcinoma can be related to methylation status of the Epstein-Barr virus BNLF-1 5′-flanking region. J. Virol. **65**:1558–1567.

36. **Marchini, A., B. Tomkinson, J. I. Cohen, and E. Kieff.** 1991. BHRF1, the Epstein-Barr virus gene with homology to Bcl2, is dispensable for B-lymphocyte transformation and virus replication. J. Virol. **65**:5991–6000.
37. **Minarovitz, J., S. Minarovitz-Kormuta, B. Ehlin-Henriksson, K. Falk, and G. Klein.** 1991. Host cell phenotype dependent methylation pattern of Epstein-Barr virus DNA. J. Gen. Virol. **75**:1591–1599.
38. **Mocarski, E. S.** 1993. Cytomegalovirus biology and replication, p. 173–226. *In* B. Roizman, R. J. Whitley, and C. Lopez (ed.), The human herpesvirus. Raven Press, New York.
39. **Morrison, J. M., H. M. Keir, H. Subak-Sharpe, and L. V. Crawford.** 1967. Nearest neighbor base sequence analysis of the deoxyribonucleic acids of a further three mammalian viruses: simian virus 40, human papilloma virus and adenovirus type 2. J. Gen. Virol. **1**:101–108.
40. **Niwa, O., Y. Yokota, H. Ishida, and T. Sugahara.** 1983. Independent mechanisms involved in suppression of the Moloney leukemia virus genome during differentiation of murine teratocarcinoma cells. Cell **32**:1105–1113.
41. **Peck, L. J., A. Nordheim, A. Rich, and J. C. Wang.** 1982. Flipping cloned $d(pCpG)_n d(pCpG)_n$ DNA sequences from right- to left-handed helical structure by salt, Co(III), or negative supercoiling. Proc. Natl. Acad. Sci. USA **79**:4560–4564.
42. **Rideout, W. M., III, G. A. Coetzee, A. F. Olumi, and P. A. Jones.** 1990. 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. Science **249**:1288–1290.
43. **Schetter, C., B. Grünemann, I. Hölker, and W. Doerfler.** 1993. Patterns of frog virus 3 DNA methylation and DNA methyltransferase activity in nuclei of infected cells. J. Virol. **67**:6973–6978.
44. **Selker, E. U.** 1990. DNA methylation and chromatin structure: a viewfrombelow.TrendsBiochem.Sci.**15**:103–107.
45. **Selker, E. U.** 1990. Premiotic instability of repeated sequences in *Neurospora crassa*. Annu. Rev. Genet. **24**:579–613.
46. **Shpaer, E. G., and J. I. Mullins.** 1990. Selection against CpG dinucleotides in lentiviral genes: a possible role of methylation in regulation of viral expression. Nucleic Acids Res. **18**:5793–5797.
47. **Simpson, R. T.** 1991. Nucleosome positioning: occurrence, mechanisms, and functional consequences. Prog. Nucleic Acid Res. Mol. Biol. **40**:143–184.
48. **Smith, D. B., and S. C. Ingles.** 1987. The mutation rate and variability of eukaryotic viruses: an analytical review. J. Gen. Virol. **68**:2729–2740.
49. **Subak-Sharpe, H., R. R. Burk, L. V. Crawford, J. M. Morrison, J. Hay, and H. M. Keir.** 1966. An approach to evolutionary relationships of mammalian DNA viruses through analysis of the pattern of nearest neighbor base sequences. Cold Spring Harbor Symp. Quant. Biol. **31**:737–748.
50. **Sutter, D., M. Westphal, and W. Doerfler.** 1978. Patterns of integration of viral DNA sequences in the genomes of adenovirus type 12-transformed hamster cells. Cell **14**:569–585.
51. **Tazi, J., and A. Bird.** 1990. Alternative chromatin structure at CpG islands. Cell **60**:909–920.
52. **Tilghman, S. M.** 1993. DNA methylation: a phoenix rises. Proc. Natl. Acad. Sci. USA **90**:8761–8762.
53. **Tooze, J.** 1982. DNA tumor viruses. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
54. **Travers, A. A., and J. W. R. Schwabe.** 1993. Spurring on transcription. Curr. Biol. **3**:898–900.
55. **Willis, D. B., and A. Granoff.** 1980. Frog virus 3 DNA is heavily methylated at CpG sequences. Virology **107**:250–257.
56. **Youssoufian, H., S. M. Hammer, M. S. Hirsch, and C. Mulder.** 1982. Methylation of the viral genome in an *in vitro* model of herpes simplex virus latency. Proc. Natl. Acad. Sci. USA **79**:2207–2210.