The dinucleotide CpG is highly underrepresented in the genome of human immunodeficiency virus type 1 (HIV-1). To identify the source of CpG depletion in the HIV-1 genome, we investigated two biological mechanisms: (1) CpG methylation-induced transcriptional silencing and (2) CpG recognition by Toll-like receptors (TLRs). We hypothesized that HIV-1 has been under selective evolutionary pressure by these mechanisms leading to the reduction of CpG in its genome. A CpG depleted genome would enable HIV-1 to avoid methylation-induced transcriptional silencing and/or to avoid recognition by TLRs that identify foreign CpG sequences. We investigated these two hypotheses by determining the sequence context dependency of CpG depletion and comparing it with that of CpG methylation and TLR recognition. We found that in both human and HIV-1 genomes the CpG motifs flanked by T/A were depleted most and those flanked by C/G were depleted least. Similarly, our analyses of human methylome data revealed that the CpG motifs flanked by T/A were methylated most and those flanked by C/G were methylated least. Given that a similar CpG depletion pattern was observed for the human genome within which CpGs are not likely to be recognized by TLRs, we argue that the main source of CpG depletion in HIV-1 is likely host-induced methylation. Analyses of CpG motifs in over 100 viruses revealed that this unique CpG representation pattern is specific to the human and simian immunodeficiency viruses.

**Key words:** CpG depletion, CpG methylation, HIV, Toll-like receptor.

**Introduction**

In the human immunodeficiency virus type 1 (HIV-1) genome, the frequency of CpG is much less than what is expected based on the frequencies of its constituent mononucleotides C and G (Shpaer and Mullins 1990; Karlin et al. 1994). Two hypotheses can be considered to explain the negative selection against CpG. The first hypothesis is that HIV-1 has evolved under a selection pressure from the host methylation machinery leading to a CpG depleted genome, which enables HIV-1 to avoid methylation-induced transcriptional silencing (van der Kuyl and Berkhout 2012). The second hypothesis is that there are host factors such as toll-like receptors (TLR) that identify and/or target CpG within foreign genomes (Bauer et al. 2001; Boehme and Compton 2004; Greenbaum et al. 2008, 2009; Chang and Altfeld 2009; Jimenez-Baranda et al. 2011; Pezda et al. 2011), thus putting pressure on HIV-1 to reduce its CpG level.

DNA methylation plays an important role in regulation of gene expression. Often methylation of CpG islands within gene promoters leads to transcriptional suppression (Jones 2012). However, contradictory data exist with regards to correlation between transcriptional suppression and promoter methylation in HIV-1 (Singh and Pauza 1992; Pion et al. 2003; Ishida et al. 2006; Blazkova et al. 2009, 2012; Duverger et al. 2009; Weber et al. 2014).

Methylation of cytosine within CpG dinucleotide significantly increases its mutation rate by spontaneous deamination to thymine (Zhang and Mathews 1994; Fryxell and Moon 2005). As such, CpG depletion of the human genome is attributed to methylation-induced deamination of C to T within CpG (Bird 1980; Simmen 2008). However, it is not clear whether this mechanism is also responsible for CpG depletion in HIV-1.

Toll-like receptors have been implicated as an alternative mechanism responsible for CpG depletion (Greenbaum et al. 2008, 2009; Jimenez-Baranda et al. 2011; Pezda et al. 2011). It is thought that the human Influenza A has been originated from an avian source (Greenbaum et al. 2008). The frequency of CpG in the influenza A genome has decreased significantly after the influenza virus crossed an interspecies barrier from avian to human, suggesting a role for host-related factors identifying CpG patterns. CpG motifs were reported to be depleted most when flanked by T and A nucleotides (Greenbaum et al. 2008, 2009; Jimenez-Baranda et al. 2011). In addition, host immune factors such as toll-like receptor 9 (TLR-9) identify unmethylated CpG within the context of T
and A in foreign DNA. Thus, this mechanism was postulated as a source of CpG depletion in influenza (Greenbaum et al. 2008, 2009; Jimenez-Baranda et al. 2011). Nevertheless, being localized in the endoplasmic reticulum (Barton et al. 2006; Chockalingam et al. 2009), TLR-9 is not likely to have access to the HIV-1 DNA, thus TLR-9 is unlikely to have played role in the depletion of CpG motifs in the HIV-1 genome.

The CpG depletion of the human genome, in addition to the reported sequence context dependency of CpG depletion in influenza virus and its possible link to components of the human immune system inspired us to explore the sequence contexts of CpG depletion and methylation in the HIV-1 genome. To investigate the mechanism responsible for CpG depletion in HIV-1, we analyzed the representation of all CpG motifs up to tetra-nucleotides in over 2,000 complete genome sequences of human and simian immunodeficiency retroviruses.

### Methods

**Data Acquisition**

We downloaded a total of 2,368 near full-length HIV-1 sequences from the Los Alamos National Laboratory database (www.hiv.lanl.gov) on October 2014. The sequences of human chromosomes were obtained from the University of California, Santa Cruz (UCSC) genome browser build 37. We obtained the whole genome methylation data of peripheral blood monocytes (PBMC) (Li et al. 2010) from Genbank, available full genome sequences of over 100 human and nonhuman viruses. Table 1 shows representative examples of viruses from all four groups of single-stranded RNA (ssRNA), double-stranded RNA (dsRNA), single-stranded DNA (ssDNA) and double-stranded DNA (dsDNA). The remaining viruses are listed in supplementary table S1, Supplementary Material online. We also downloaded influenza sequences (1,000 human influenza type A subtype H1N1, 800 human influenza type A subtype H3N2, and 200 human influenza type B) from the Influenza Research Database (IRD), (www.fludb.org/irdb/home.jsp?decorator=influenza; last accessed April 2015).

### Data Analysis

**Analysis of CpG Representation**

We define "representation (D-ratio)" as the ratio of observed frequency ($P_{\text{obs}}$) of a motif over its expected frequency ($P_{\text{exp}}$) in the genome. The $P_{\text{obs}}$ of a motif is defined as the number of times that motif appears in the sequence divided by the total number of all motifs with the same length (Burge et al. 1992; Leung et al. 1996). The $P_{\text{exp}}$ can be calculated in different ways.

\[
P_{\text{exp}}(\text{CpG}) = P_{\text{obs}}(\text{C}) \times P_{\text{obs}}(\text{G}) \times P_{\text{obs}}(\text{T})
\]

\[
P_{\text{exp}}(\text{GpT}) = P_{\text{obs}}(\text{G}) \times P_{\text{obs}}(\text{T})
\]

\[
P_{\text{exp}}(\text{CpT}) = P_{\text{obs}}(\text{C}) \times P_{\text{obs}}(\text{T})
\]

### Table 1. List of Representative Viruses Investigated in This Study.

<table>
<thead>
<tr>
<th>Virus Type</th>
<th>Host</th>
<th>Genome Length (nt)</th>
<th>Number of Sequences</th>
<th>CpG D-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV-1 (retrovirus)*</td>
<td>ssRNA</td>
<td>Human</td>
<td>9,000</td>
<td>2,368</td>
</tr>
<tr>
<td>HIV-2 (retrovirus)*</td>
<td>ssRNA</td>
<td>Human</td>
<td>9,000</td>
<td>20</td>
</tr>
<tr>
<td>SIVcpz (retrovirus)*</td>
<td>ssRNA</td>
<td>Simian</td>
<td>9,000</td>
<td>29</td>
</tr>
<tr>
<td>SIVsm (retrovirus)*</td>
<td>ssRNA</td>
<td>Simian</td>
<td>9,300</td>
<td>5</td>
</tr>
<tr>
<td>HTLV-1 (retrovirus)*</td>
<td>ssRNA</td>
<td>Human</td>
<td>9,000</td>
<td>20</td>
</tr>
<tr>
<td>SFV (retrovirus)*</td>
<td>ssRNA</td>
<td>Simian</td>
<td>12,900</td>
<td>28</td>
</tr>
<tr>
<td>HERV-K (retrovirus)*</td>
<td>ssRNA</td>
<td>Human</td>
<td>9,300</td>
<td>22</td>
</tr>
<tr>
<td>SINV</td>
<td>ssRNA</td>
<td>Human</td>
<td>11,700</td>
<td>3</td>
</tr>
<tr>
<td>Semliki forest</td>
<td>ssRNA</td>
<td>Human</td>
<td>11,440</td>
<td>3</td>
</tr>
<tr>
<td>Influenza A H1N1</td>
<td>ssRNA</td>
<td>Human</td>
<td>12,500</td>
<td>1,000</td>
</tr>
<tr>
<td>HCV</td>
<td>ssRNA</td>
<td>Human</td>
<td>9,600</td>
<td>6</td>
</tr>
<tr>
<td>CTFV (RNA virus)</td>
<td>dsRNA</td>
<td>Human</td>
<td>25,200</td>
<td>11</td>
</tr>
<tr>
<td>RV Type C (RNA virus)</td>
<td>dsRNA</td>
<td>Human</td>
<td>18,500</td>
<td>2</td>
</tr>
<tr>
<td>AAV</td>
<td>ssDNA</td>
<td>Human</td>
<td>4,700</td>
<td>11</td>
</tr>
<tr>
<td>BoV</td>
<td>ssDNA</td>
<td>Human</td>
<td>5,300</td>
<td>6</td>
</tr>
<tr>
<td>JC polyomavirus</td>
<td>dsDNA</td>
<td>Human</td>
<td>5,120</td>
<td>40</td>
</tr>
<tr>
<td>HBV</td>
<td>dsDNA</td>
<td>Human</td>
<td>3,300</td>
<td>400</td>
</tr>
</tbody>
</table>

**Note.—** A full list of viruses studied is given in supplementary table S1, Supplementary Material online.

*—Retroviruses package single-stranded RNA in their particles but their life cycle also includes a DNA intermediate.

HIV, human immunodeficiency virus; SIVcpz, Simian immunodeficiency virus (chimpanzee); SIVsm, Simian immunodeficiency virus (Sooty Mangabey); HTLV, human T-lymphotropic virus; SFV, Simian foamy virus; HERVK, human endogenous retrovirus type K; SINV, Sindbis virus; HCV, hepatitis C virus; CTFV, Colorado tick fever virus; RV, rotavirus; AAV, Adeno-associated virus; BoV, Bocavirus; HBV, hepatitis B virus.
models of conditional probabilities (Ebrahimi et al. 2012; Anwar et al. 2013). In this method (Anwar et al. 2013), the expected frequency of a motif is estimated using the observed frequencies of the motif constituents, considering the overlapping nucleotide(s). Examples of first and second order Markov models are given in Equations (4) and (5).

\[
P_{\text{exp}}(\text{CpGpT}) = \frac{P_{\text{obs}}(\text{CpG}) \times P_{\text{obs}}(\text{GpT})}{P_{\text{obs}}(\text{G})}
\]

\[
P_{\text{exp}}(\text{CpGpTpA}) = \frac{P_{\text{obs}}(\text{CpGpT}) \times P_{\text{obs}}(\text{GpTpA})}{P_{\text{obs}}(\text{GpT})}
\]

In this study, we used first and second order models to calculate the expected frequencies of tri-nucleotides and tetra-nucleotides, respectively. We then calculated, for each motif, its representation (D-ratio) by dividing the \(P_{\text{obs}}\) of the motif by its \(P_{\text{exp}}\). To compare the frequencies of different motifs, we performed nonparametric Mann–Whitney tests. To determine the statistical significance of correlations, we used nonparametric Spearman correlation tests.

**Analysis of %Methylation of CpG**

We performed three different analyses to quantify the percentage of methylated CpG tri- and tetra-nucleotides using the whole genome PBMC DNA methylene data (Li et al. 2010) from the NGSmethDB database (http://bioinfo2.ugr.es:8080/NGSmethDB/; last accessed June 2015) (Hackenberg et al. 2011; Geisen et al. 2014). In the first analysis, we quantified %methylation in whole chromosomes 1, 2, 3 and 10. For the second and third analyses, we quantified %methylation inside and outside CpG islands, respectively. For these two analyses, we only used data from chromosomes 1 and 2. It is worth noting that analysis of data from a single chromosome at a time returned the same results, implying that the reported data in this article is independent of which chromosome(s) is(are) used. For each analysis, we report %methylation as an average of all analyzed chromosomes. To identify CpG islands, we used the Matlab function “cpgisland” with a 100-bp moving window, which is a default setting of this function. The CpG islands were defined as regions within which GC content is >50% and the ratio of the observed over expected CpG content (based on the frequencies of C and G) is >60%.

**Results**

Analysis of the representation of CpG motifs shown in figure 1A–D indicates that CpG depletion is sequence context dependent. The CpG tri-nucleotides flanked by T or A (shown by circles) are significantly less represented than those flanked by C or G (shown by triangles) in both HIV-1 (fig. 1A) \((P = 0.03, \text{Mann–Whitney})\) and human (fig. 1C) \((P = 0.03)\). The same pattern is observed for CpG tetra-nucleotides shown in figure 1B and D. The CpG tetra-nucleotides having T and/or A only (e.g., CpGpTpA and ApCpGpA) were significantly less represented than CpG motifs flanked by C and/or G only (CpGpCpG and CpGpCpG) \((P = 0.04\) and 0.0002 for HIV-1 and human genomes, respectively).

To show that the observed CpG representation pattern of the HIV-1 genome is not an artifact of HIV-1 genome biases such as A-richness and possible hypermutation signatures of APOBEC3 enzymes, we performed the following two analyses:

a. We performed a representation analysis for the motif GpC, which has the same composition as CpG but is not under methylation-induced mutation pressure. If our reported CpG depletion pattern (i.e., greater depletion of CpG motifs flanked by T/A compared with those flanked by C/G) is due to A-richness of the HIV-1 genome, one would expect to see the same pattern for GpC. Our analyses of GpC tri- and tetra-nucleotides did not show the pattern that we observed for CpG (supplementary fig. S1a and b, Supplementary Material online).

b. We performed two separate representation analyses, one for hypermutated HIV-1 sequences and one for nonhypermutated HIV-1 sequences. As predicted, the CpG representation patterns of nonhypermuted and hypermutated sequences were not different (supplementary fig. S2a and b, Supplementary Material online).

These two analyses indicate that genome composition biases such as A-richness and APOBEC-induced hypermutation signature do not affect our analysis.

We also investigated CpG representation in the HIV-1 long terminal repeats (LTRs), which are not transcribed. The results showed that similar to the rest of the HIV-1 genome, in the HIV-1 LTR regions, CpG motifs flanked by T/A are less represented than those flanked by C/G (supplementary fig. S3a and b, Supplementary Material online). To investigate whether codon bias has played role in the observed lower representation of CpG motifs flanked by T/A, we compared the codon usage of Arginine, which is coded by all four possible CpG-containing codons, namely CGA, CGU, CGC, and CGG (Pandit and Sinha 2011). If the observed CpG depletion pattern (lower representation of CGT and CGA compared with those of CGC and CGG) is due to the HIV-1 codon bias, one would expect to see a lower usage, for Arginine, of the codons CGU and CGA compared with CGC and CGG. Our results showed that in the HIV-1 genes codon CGA is used the most (see supplementary fig. S4, Supplementary Material online). This suggests that CpG depletion pattern that we report here is not an artifact of codon bias of the HIV-1 genes.

To investigate whether the observed pattern of CpG depletion is specific to HIV-1 or is common among viruses, we analyzed full genome sequences from a wide range of viruses including viruses with ssDNA, dsDNA, ssRNA, and dsRNA genomes having a wide range of CpG representations (see table 1 and supplementary table S1, Supplementary Material online). The results of few viruses from each group are shown in figure 2 and those of additional 103 viruses investigated are given in supplementary figure S5a–j, Supplementary Material online. Similar to HIV-1, in HIV-2 and in their simian immuno-nodeficiency counterpart viruses (i.e., SIVcpz and SIVsm) the representation of CpG was lower when it was flanked by T/A nucleotides. All other SIV sequences studied (e.g., SIVagm, SIVste) showed a similar pattern as that of HIV-1.
SIVMND and SIVMUS also showed the same CpG representation pattern. The only exception was SIVGOR (see supplementary fig. S5a, Supplementary Material online). This pattern was not observed in any of other viruses even in the human endogenous retroviruses HERV-K (see supplementary fig. S5b–j, Supplementary Material online). Contrary to what has been reported previously (Greenbaum et al. 2008; Jimenez-Baranda et al. 2011), we found that in the influenza virus, CpG depletion is not sequence context dependent (fig. 2 and supplementary fig. S5b and h, Supplementary Material online).

The lifecycle of HIV-1 includes both RNA and DNA stages. To investigate whether the mechanism responsible for CpG depletion in HIV-1 has acted on the HIV-1 RNA genome or on the integrated HIV-1 DNA, we compared the representation of each CpG motif with that of its reverse complement motif (e.g., CpGpA with TpCpG) as shown in figure 3. If the mechanism has acted at a DNA level, and assuming there has not been a strand targeting bias, we expect to see reverse complementary motifs represented equally in the HIV-1 genome. As displayed in figure 3A, there is a positive correlation between the representations of reverse complementary tri-nucleotide CpG motifs. Strikingly, this pattern is also present for tetra-nucleotide CpG motifs (fig. 3B). These data suggest that the mechanism responsible for context-dependent CpG depletion has acted on the HIV-1 DNA.

In order to investigate whether methylation has played a role in the CpG depletion of HIV-1 genome, we analyzed DNA methylation patterns in human PBMC. The results of the percentage methylation of CpG tri- and tetra-nucleotide motifs in the human genome are shown in figure 4A and B, respectively. The results indicate that CpG methylation is a sequence context dependent mechanism. The CpG motifs flanked by T/A are methylated at a higher rate compared with those flanked by C/G. To ensure that the observed pattern is a general feature of methylation and is not affected by clustering of CpGs in the human genome, we performed two additional analyses in which only CpGs inside CpG islands or outside CpG islands were included. The results of the analysis of CpGs inside CpG islands are shown in figure 4C and D for tri- and tetra-nucleotides, respectively. The results of the
analysis of CpGs falling outside CpG islands are given in figure 4E and F for tri- and tetra-nucleotides, respectively. As indicated, regardless of the positioning of CpG sites inside or outside CpG islands, CpG motifs flanked by T/A exhibit a higher percentage of methylation. These data suggest that the sequence context dependency of methylation is independent of CpG distribution in the human genome.

We also computed the correlation between %methylation and depletion of CpG tri- and tetra-nucleotide motifs. Figure 5A and B show a significant inverse correlation between CpG methylation and CpG depletion, confirming that CpG motifs that are methylated most are underrepresented most, and vice versa.

In figure 5B, all CpG tetra-nucleotide motifs, namely NNCG, CGNN, and NCGN (N: C/G or T/A) were included. We also performed two separate analyses for NCGN and (CGNN and NNCG) motifs. The results, presented in supplementary figure 56, Supplementary Material online, show a significant inverse correlation between %methylation and representation of NNCG and CGNN motifs in both human and HIV-1 genomes. However, for the motifs NCGN, the inverse correlation was insignificant at 95% confidence level in both human and HIV-1. Additionally, we compared the representation of CpG motifs flanked by T/A and those flanked by C/G. The results revealed that, in both human and HIV-1, only for NNCG and CGNN motifs but not for NCGN motifs, the representation is lower when CpG is flanked by T/A (see supplementary fig. S7, Supplementary Material online).

Putting together, these data indicate that the methylation rate of CGNN and NNCG is higher when N is T/A compared with when N is C/G. As a result the representation of CGNN and NNCG is lower when N is T/A compared with when N is.
C/G. Importantly, this pattern is present in both human and HIV-1 genomes. In contrast, the methylation rate and representation of NCGN motifs appears to be independent of the type of flanking nucleotides. Again, this pattern is present in both human and HIV-1 genomes. These data suggest: (1) methylation of CpG sites is nonrandom and sequence context dependent and (2) The mechanism responsible for CpG depletion of the human and HIV-1 genomes is likely the same.

Discussion

CpG is highly underrepresented in the genome of HIV-1 and many other viruses (Kypr et al. 1989; Burge et al. 1992; Karlin et al. 1994; van der Kuyl and Berkhout 2012; Cheng et al. 2013). Depletion of CpG is also a known feature of the human and other vertebrate genomes. CpG depletion in vertebrates is thought to be a result of CpG methylation followed by
spontaneous deamination to TpG on the same strand or a mismatch repair to CpA on the complementary strand (Bird 1980; Simmen 2008). However, it is not clear why CpG is depleted in the genome of viruses such as HIV-1. To answer this question, we quantified the representation of CpG tri- and tetra-nucleotide motifs in a diverse population of full genome HIV-1 sequences and also in the human genome. We then compared the CpG depletion patterns with the CpG methylation and TLR recognition patterns to identify the source of CpG depletion in HIV-1. Analysis of the representation of CpG motifs revealed that in both human and HIV-1 genomes the motifs flanked by T/A have been depleted more compared with those flanked by C/G (fig. 1). These data suggest that the mechanism responsible for CpG depletion has mutated CpG motifs in a sequence context dependent manner in both genomes.

The host defense molecule TLR-9 (Hemmi et al. 2000; Barton 2007) recognizes foreign CpG DNA, mainly in the context of T and A (Bauer et al. 2001; Greenbaum et al. 2008, 2009; Jimenez-Baranda et al. 2011; Pezda et al. 2011). It has been speculated that human influenza virus had evolved to reduce, in its genome, the level of CpG flanked by T/A to avoid recognition by TLR-9 (Bauer et al. 2001; Greenbaum et al. 2008, 2009; Jimenez-Baranda et al. 2011; Pezda et al. 2011). Although we were inspired by the results reported previously for the influenza virus, our representation analyses of whole genome and different segments of influenza virus did not show the reported patterns. Figure 2 and supplementary figure S5b and h, Supplementary Material online, show that there is no difference between the representation of CpG motifs flanked by T/A and C/G in the genome of human and avian influenza viruses. Nevertheless, given that the human genome exhibits the same CpG depletion pattern as HIV-1 and that the human genome is unlikely to have been affected by TLR-9 (Bauer et al. 2006), this suggests TLR-9 recognition may not be the source of CpG depletion in HIV-1.

Our results shown in figure 3 indicate that CpG reverse complementary motifs are nearly identically represented in the HIV-1 genome suggesting that CpG depletion has occurred at the DNA stage of the HIV-1 life cycle, likely in HIV-1 proviruses. Additionally, we note that TLR-9 exist in the endolysosomal compartments of innate immune cells (Barton et al. 2006; Chockalingam et al. 2009), thus, its access to the HIV-1 DNA during reverse transcription in the cytoplasm or after the HIV-1 DNA is integrated into the human genome, seems unlikely.

In addition to TLR, in this study, we investigate methylation-induced mutation as a potential source of CpG depletion in HIV-1. We reason that if methylation-induced silencing (and/or mutation) is responsible for the depletion of CpG motifs in HIV-1, those motifs that are methylated more are expected to be depleted more. In other words, we would expect CpG motifs flanked by T/A to be methylated more frequently compared with CpG motifs flanked by C/G. Figure 4 shows that the presence of T and A in the flanking regions of CpG significantly increases the rate of methylation. In addition, methylation of double-stranded DNA is expected to act equally on a motif and its reverse complement, which is what we observe in figure 3. The inverse correlation between %methylation and representation of CpG motifs (fig. 5) is another piece of evidence suggesting that methylation is the source of CpG depletion in HIV-1. The HIV-1 genome is only 9000 bp in size. In such a small genome, it is striking that even CpG tetrانucleotides show a pattern that points to methylation as the source of CpG depletion in HIV-1 (figs. 1B, 3B, and 5B).

The genome of many viruses including those studied here (e.g., Influenza, HERV-K, HBV, SFV, JC Polyomavirus) is CpG depleted. However, our results indicate that among the viruses studied only immunodeficiency viruses HIV and SIV have a CpG depletion pattern in their genomes that mimics the pattern of human methylation-induced mutations. Surprisingly, even viruses such as HERV-K, which is endogenous to the human genome and JC Polyomavirus, which has a dramatically reduced CpG level (see table 1), do not show a CpG depletion pattern that can be attributed to methylation-induced mutation (fig. 2 and supplementary fig. S5a–j, Supplementary Material online). The exclusive host CpG-mimicry in HIV and SIV may provide a unique advantage in these viruses; nevertheless further evidence is needed to support a viral adaptation model. Our study provides evidence
that suggest CpG methylation is the source of CpG depletion in the HIV and SIV genomes. However, what drives CpG depletion in other viruses, including those infecting hosts that are not CpG depleted remains to be understood.

**Supplementary Material**

Supplementary figures S1–S7 and tables S1 and S2 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Author Contributions**


**References**


