CpG dinucleotides inhibit HIV-1 replication through zinc finger antiviral protein (ZAP)-dependent and -independent mechanisms

Running Head: Position-dependent effect of CpG dinucleotides.

Mattia Ficarelli,a Irati Antzin-Anduetza,a Rupert Hugh-White,b Andrew E. Firth,c Helin Sertkaya,a Harry Wilson,a Stuart JD Neil,a Reiner Schulz,b Chad M Swanson a#

a: Department of Infectious Diseases, King's College London, London, United Kingdom
b: Department of Medical and Molecular Genetics, King's College London, London, United Kingdom
c: Division of Virology, University of Cambridge, Cambridge, United Kingdom

#: Address correspondence to Chad M Swanson, chad.swanson@kcl.ac.uk

M.F. and I.A.A. contributed equally to this work.

Abstract word count: 201
Text word count: 7688
ABSTRACT

CpG dinucleotides are suppressed in the genomes of many vertebrate RNA viruses, including HIV-1. The cellular antiviral protein ZAP binds CpGs and inhibits HIV-1 replication when they are introduced in the viral genome. However, it is not known if ZAP-mediated restriction is the only mechanism driving CpG suppression. To determine how CpG dinucleotides affect HIV-1 replication, we increased their abundance in multiple regions of the viral genome and analyzed the effect on RNA expression, protein abundance and infectious virus production. We found that the antiviral effect of CpGs does not correlate with their abundance. Interestingly, CpGs inserted into some regions of the genome sensitize the virus to ZAP antiviral activity more efficiently than other regions and this sensitivity can be modulated by interferon treatment or ZAP overexpression. Furthermore, the sensitivity of the virus to endogenous ZAP correlated with its sensitivity to the ZAP cofactor KHNYN. Finally, we show that CpGs in some contexts can also inhibit HIV-1 replication by ZAP-independent mechanisms and one of these is the activation of a cryptic splice site at the expense of a canonical splice site. Overall, we show that the location and sequence context of the CpG in the viral genome determines its antiviral activity.

IMPORTANCE

Some RNA virus genomes are suppressed in the nucleotide combination of a cytosine followed by a guanosine (CpG), indicating that they are detrimental for the virus. The antiviral protein ZAP binds viral RNA containing CpGs and prevents the virus from multiplying. However, it remains unknown how the number and position of CpGs in viral genomes affect restriction by ZAP and whether CpGs have other
antiviral mechanisms. Importantly, manipulating the CpG content in viral genomes could help create new vaccines. HIV-1 shows marked CpG suppression and, by introducing CpGs into its genome, we show that ZAP efficiently targets a specific region of the viral genome, the number of CpGs does not predict the magnitude of antiviral activity and that CpGs can inhibit HIV-1 gene expression through a ZAP-independent mechanism. Overall, the position of CpGs in the HIV-1 genome determines the magnitude and mechanism through which they inhibit the virus.
INTRODUCTION

The frequency of CpG dinucleotides is suppressed in many vertebrate RNA viruses, indicating that they may be deleterious (1-5). Supporting this hypothesis, increasing the CpG content in picornaviruses and influenza A virus inhibits their replication (6-10). However, the mechanisms by which CpG dinucleotides attenuate viral replication remain unclear. Importantly, introduction of CpG dinucleotides into viral genomes using synthetic biology techniques may be a new way to develop live attenuated virus vaccines and a full understanding of how CpG dinucleotides inhibit viral replication is necessary to develop this approach (9, 10).

HIV-1 encodes the three polyproteins found in all retroviruses (Gag, Pol and Env), two regulatory proteins (Tat and Rev) and four accessory proteins (Vif, Vpr, Vpu and Nef) (11). CpG dinucleotides are suppressed throughout the HIV-1 genomic RNA (gRNA) and introducing CpGs into gag or env inhibits viral replication (12-16). Furthermore, analysis of clinical HIV-1 samples found that mutations that create new CpG dinucleotides in HIV-1 are twice as costly as those that do not and increased CpG dinucleotide abundance in env may predict disease progression (17, 18). There are at least four mechanisms by which CpG dinucleotides could inhibit HIV-1. First, CpG DNA methylation-induced transcriptional silencing could repress viral gene expression (12, 13, 19). Second, introduction of CpGs into cis-acting elements or structures required for viral replication may render them non-functional. Third, CpGs could create deleterious cis-acting elements or structures. Fourth, they could act as a pathogen-associated molecular pattern (PAMP) that is recognized by the innate immune system. Supporting the hypothesis that CpGs in viral RNA could be a PAMP, the antiviral protein ZAP has recently been shown to bind regions of HIV-1 RNA.
containing CpG dinucleotides and inhibit HIV-1 with increased CpG content in env (16).

ZAP (encoded by the gene ZC3HAV1) was initially discovered as a cellular factor inhibiting murine leukemia virus (MLV) gene expression (20). Subsequent studies have shown that ZAP also inhibits alphaviruses, filoviruses and hepadnaviruses as well as some flaviviruses and picornaviruses (21-25). ZAP inhibits viral replication by binding viral RNA and targeting it for degradation and/or inhibiting its translation (20, 21, 26-28). It may also have other mechanisms of antiviral activity. However, ZAP does not restrict all viruses and yellow fever virus, Zika virus, dengue virus, herpes simplex 1 virus, vesicular stomatitis virus and poliovirus are resistant to its antiviral activity (21, 24). ZAP does not have enzymatic activity and interacts with other cellular proteins, such as TRIM25 and KHNYN, to restrict viral replication (29-31).

Why some viruses are sensitive to ZAP and others are resistant, and whether the CpG abundance and context in viruses determines this, remains unknown. It is also unclear whether the deleterious effect of CpG dinucleotides on viral replication is mediated exclusively by ZAP or through additional mechanisms.

Overall, the specific mechanisms by which CpG dinucleotides inhibit viral replication are not well understood for any virus. Because CpGs are highly suppressed in HIV-1, it is an excellent model virus to study the antiviral effects of this dinucleotide. In this study, we introduced CpGs in different contexts and regions of the viral genome and analyzed how they restrict viral gene expression and replication. First, we determined whether there was a position-dependent effect of CpGs on viral replication. Second, we analyzed whether there was a correlation between endogenous ZAP activity and
the abundance of CpG dinucleotides. Third, we tested whether increasing ZAP abundance increased its antiviral effect. In sum, we found that CpGs in different contexts and locations inhibited viral gene expression through ZAP-dependent inhibition of gene expression and ZAP-independent changes in pre-mRNA splicing. Importantly, the number of introduced CpG dinucleotides did not predict the magnitude of their antiviral activity or the inhibition by endogenous ZAP. ZAP appears to target a specific region in env containing introduced CpGs more efficiently than other regions in the viral genome, though high levels of ZAP can target most regions of the genome containing CpGs. Our results indicate that the context and position of CpG dinucleotides in the HIV-1 genome determines how they inhibit viral replication through ZAP-dependent and -independent mechanisms.

RESULTS
To determine how CpG suppression in HIV-1 compared with other viruses that infect vertebrates, we compared the CpG frequency in a panel of viruses (Fig. 1A, Data set S1). Because RNA viruses have large variations in the frequency of A, C, G and U in their genomes, we calculated the CpGs/kilobase (kb) of RNA and the observed/expected ratio of CpGs. This analysis showed that there is a broad range of CpG suppression in RNA viruses. As previously shown, togaviruses show little CpG suppression with >40 CpGs/kb and an observed/expected ratio of >0.75 (1-5). Many viruses show moderate CpG suppression with an observed/expected CpG ratio of ~0.5. However, there are some viruses in which CpG abundance is highly suppressed including hepatitis A virus, respiratory syncytial virus and HIV-1 (Fig. 1A, Data set S1). Within the retrovirus family, lentiviruses have high levels of CpG suppression (6-23 CpGs/kb, 0.2-0.4 observed/expected) and alpharetroviruses have
low levels of suppression (~50 CpGs/kb, ~0.7 observed/expected, Fig. 1B, Data set S1). Viruses highly related to HIV-1 have ~10 CpGs/kb and an ~0.2 observed/expected ratio.

In addition to CpG dinucleotides, UpA dinucleotides are suppressed in many RNA viruses (1, 3, 4). Recently, it was reported that ZAP interacts with viral RNA containing UpAs and restricts echovirus 7 containing introduced UpAs (32). Therefore, we analyzed the UpA abundance in our panel of viruses (Fig. 1C, Data set S1). While some vertebrate RNA viruses, such as flaviviruses, show marked UpA suppression (<30 UpAs/kb and <0.5 observed/expected, Fig. 1C), retroviruses are not substantially suppressed in UpA frequency (Fig. 1D). Specifically, viruses closely related to HIV-1 have ~70 UpAs/kb and ~0.9 observed/expected. In sum, CpGs but not UpAs appear to be potently suppressed in HIV-1.

To better understand the potential evolutionary pressures that have led to CpG suppression in HIV-1, we explored the effect of introducing CpGs into different regions of the viral genome using synonymous mutations. However, HIV-1 contains several overlapping reading frames that constrain where CpGs can be introduced (Fig. 2A). Furthermore, it is important that the synonymous mutations introducing CpGs do not disrupt RNA elements that regulate viral replication. The HIV-1 open reading frames contain multiple cis-acting regulatory elements including the programmed ribosomal frameshift sequence in gag (33), the Rev-response element (RRE) in env (34), polyuridine tracts in pol and nef (35) and splicing signals in pol, vif, vpr, tat, rev and env (36). In addition, there could be uncharacterized elements. Therefore, we identified sequences in the HIV-1 open reading frames that contain
reduced variability at synonymous sites, which could indicate the presence of functional RNA elements (37). This analysis identified many of the known HIV-1 linear and structural RNA regulatory elements including the region at the 5’ end of *gag* required for dimerization and encapsidation, the ribosomal frameshift sequence required for Pol translation, the polypurine tracts required for reverse transcription, several splicing regulatory sequences and the RRE (Fig. 2B-E and Data set S2). We have synonymously introduced CpG dinucleotides into *gag*, *pol* and *env* sequences that the analysis revealed were unlikely to contain *cis*-acting elements (Data set S2).

**The number of CpGs introduced into the HIV-1 genome does not correlate with the antiviral effect or ZAP-sensitivity.**

An important experimental consideration when studying how CpGs regulate HIV-1 is that CpG DNA methylation-induced transcriptional silencing could potentially inhibit HIV-1 gene expression. However, CpGs in plasmids amplified in bacteria are not methylated when the plasmid is transiently transfected into mammalian cells (38-40) and we have therefore used an experimental approach in which HIV-1 proviral DNA plasmids are transfected into HeLa or 293T cells. A region in *env* immediately after *vpu* does not contain any detectable *cis*-acting RNA elements, which makes it a good region to analyze the effect of introducing CpGs (Fig. 2 and Data set S2). It has previously been shown that introducing 36 CpGs into *env* nucleotides (nt) 86-561 (HIV-1*env*86-561CpG) inhibits HIV-1 genomic RNA abundance, Gag expression, Env expression, and infectious virus production (Fig. 3A, 3C and 3E, Table 1) (16, 31).

Importantly, this restriction is eliminated in ZAP knockout cells (Figure 3B, 3C and 3E), indicating that it is due to the recognition of the CpG dinucleotides by ZAP. To further characterize how CpGs affect HIV-1 replication, we inserted 48 CpGs into *env*...
nt 611-1014 (HIV-1<em>env</em>611-1014CpG, Fig. 3A, Table 1) and analyzed their effect on HIV-1 genomic RNA abundance, Env expression, Gag expression and infectious virus production. These CpGs inhibited HIV-1 infectious virus production more potently than the 36 CpGs in HIV-1<em>env</em>86-561CpG in the control CRISPR cells (Fig. 3C). However, in the ZAP knockout cells, Gag expression, Env expression and virion production were only partially increased for HIV-1<em>env</em>611-1014CpG and infectious virus production was rescued only ~2-fold (Fig. 3C). This indicates that the 48 CpGs introduced into this region of <em>env</em> caused both ZAP-dependent and ZAP-independent suppression of infectious virus production. Nef is expressed from fully spliced mRNAs that do not contain the <em>env</em> region with the introduced CpGs (36). As expected, the CpGs introduced in HIV-1<em>env</em>86-561CpG did not decrease Nef expression (Fig. 3C). However, there was decreased Nef expression for HIV-1<em>env</em>611-1014CpG and this was not rescued in the ZAP CRISPR cells. This suggests that another mechanism, such as altered splicing, contributes to the CpG-mediated decrease in infectious virus production for HIV-1<em>env</em>611-1014CpG. Furthermore, genomic RNA abundance for HIV-1<em>env</em>611-1014CpG was not fully restored to wild type levels in ZAP knockout cells, indicating that the CpGs in HIV-1<em>env</em>611-1014CpG inhibit genomic RNA abundance through ZAP-dependent and ZAP-independent mechanisms (Fig. 3E).

We also combined the two regions in <em>env</em> containing 36 and 48 CpGs for a total of 84 CpGs (HIV-1<em>env</em>86-1014CpG, Fig. 3A, Table 1). This inhibited HIV-1 infectious virus production in an approximately additive manner compared to the two regions’ individual effects through both ZAP-dependent and ZAP-independent effects on genomic RNA abundance and viral protein expression (Fig. 3C and 3E). To determine the contribution of decreased Env expression to CpG-mediated inhibition,
we pseudotyped the viruses with the vesicular stomatitis virus glycoprotein (VSV-G) and found that the ZAP-dependent- and -independent defects in infectious virus production were still present (Fig. 3D).

We then introduced 53 CpGs into a region of pol that did not contain any known or detectable cis-acting elements (HIV-1_{pol795-1386}CpG, Fig. 2, Data set S2, Fig. 4A, Table 1). Surprisingly, this large number of CpGs caused only a small (~2-fold) reduction in Gag expression and infectious virus production in control CRISPR cells and this effect was eliminated in the ZAP knockout cells (Fig. 4B). This suggests that the magnitude of ZAP-dependent restriction is not simply proportional to the absolute number of CpGs added to the viral genome.

We have recently identified KHNYN as an essential ZAP cofactor for CpGs to inhibit HIV-1_{env86-561}CpG gene expression and infectious virus production (31). Our previous work showed that KHNYN overexpression inhibited HIV-1_{env86-561}CpG much more potently than wild type HIV-1, indicating that the introduced CpGs were required for KHNYN to inhibit HIV-1 (31). To determine if inhibition by KHNYN correlated with CpG-abundance or ZAP-sensitivity, we overexpressed KHNYN on wild type HIV-1, HIV-1_{env86-561}CpG or HIV-1_{pol795-1386}CpG (Fig. 4C-D). KHNYN antiviral activity correlated with the sensitivity of the virus to endogenous ZAP, with HIV-1_{env86-561}CpG inhibited much more potently than HIV-1_{pol795-1386}CpG or wild type HIV-1. We also tested whether endogenous KHNYN was required to restrict HIV-1_{env611-1014}CpG, HIV-1_{env86-1014}CpG and HIV-1_{pol795-1386}CpG using the KHNYN CRISPR cells that we previously characterized (31). Wild type HIV-1 infectious virus production was not affected by depleting KHNYN while HIV-1_{env86-561}CpG infectious virus production and
gene expression were substantially increased (31) (Fig. 4E). The CpG-mediated inhibition of HIV-1\textsubscript{env611-1014}CpG and HIV-1\textsubscript{env686-1014}CpG infectious virus production were partially rescued in KHNYN CRISPR cells, which correlated with their restriction by ZAP (Fig. 4E and 3C). The small decrease in infectious virus production for HIV-1\textsubscript{pol795-1386}CpG was rescued in the KHNYN CRISPR cells, indicating that the 53 introduced CpGs in \textit{pol} moderately inhibit HIV-1 through ZAP and KHNYN (Fig. 4E).

HIV-1 containing reporter genes such as green fluorescent protein (GFP) or luciferase are important experimental tools and have large numbers of CpGs in the reporter gene. Therefore, it was important to determine whether ZAP inhibits these viruses because this could confound the interpretation of results obtained under some experimental conditions. Specifically, we analyzed whether ZAP could restrict HIV-1 containing the encephalomyocarditis virus (EMCV) internal ribosome entry site (IRES) followed by eGFP (IRES-GFP) or \textit{Renilla} luciferase (IRES-\textit{Renilla}) (Fig. 5A, Table 1) (41, 42). Both IRES-GFP and IRES-\textit{Renilla} introduced 96 CpGs into the viral genome. While both reporter viruses produced less Gag and infectious virus than wild type HIV-1, ZAP depletion did not increase this (Fig. 5B). We also analyzed the effect of a Venus fluorescent protein plus linker sequence that introduces 64 CpGs as a fusion protein with Gag (Fig 5A, Table 1) (43). Interestingly, this sequence did not sensitize Gag abundance or virus-like particle (VLP) production to ZAP (Fig. 5C). Because a ribosome could displace ZAP bound to CpGs in an open reading frame as it moves along the mRNA, it is possible that CpGs in a 3’ untranslated region (UTR) could inhibit HIV-1 gene expression in a ZAP-dependent manner more effectively than CpGs in a coding region. Therefore, we inserted stop codons between Gag and Venus in both the Gag and Pol reading frames (Fig. 5A, Table 1).
However, the CpGs in the context of the 3' UTR also did not promote ZAP-mediated inhibition of Gag expression or VLP production (Fig. 5C). In sum, the total number of CpGs introduced into the HIV-1 genome does not correlate with their antiviral activity in the context of endogenous ZAP levels in HeLa cells.

Increasing ZAP abundance inhibits CpG-containing HIV-1.

We then analyzed whether increasing ZAP abundance further inhibited HIV-1 containing introduced CpGs. First, we treated control and ZAP knockout HEK293T cells with type I interferon (IFN-I) (Fig. 6). Similar to previous reports, this consistently increased ZAP-S expression by ~2-fold and had no substantial effect on ZAP-L expression (Fig. 6C) (44, 45). IFN-I inhibited all of the viruses. For wild type HIV-1, the magnitude of inhibition was similar in the control and ZAP knockout cells (Fig. 6A), which is consistent with the observations that endogenous ZAP does not target wild type HIV-1 (16, 31). The magnitude of IFN-I inhibition for HIV-1_{env66-561}CpG was reduced when ZAP was depleted, indicating that ZAP contributes to the antiviral effect of IFN-I on this virus. Importantly, IFN-I treatment augmented the ZAP-dependent inhibition of HIV-1_{pol795-1386}CpG. This suggests that CpG-rich sequences in an HIV open reading frame that are only weakly restricted by endogenous ZAP levels can be further sensitized by IFN-I. However, even in the presence of IFN-I, the magnitude of inhibition by the 53 introduced CpGs in pol was lower than the inhibition mediated by the 36 introduced CpGs in env. For HIV-1-IRES-GFP and HIV-1-IRES-

Renilla, the magnitude of IFN-I inhibition was not decreased upon ZAP depletion (Fig. 6B). This indicates that endogenous ZAP does not target these viruses, despite the large numbers of CpGs that have been introduced in the viral genome.
Because IFN-I only moderately upregulated ZAP-S, we also overexpressed ZAP-S or ZAP-L in HeLa cells. This increased ZAP-S expression ~5-fold and ZAP-L expression ~20-fold (Fig. 7C). Both ZAP isoforms inhibited all of the viruses tested at least 2-fold (Fig. 7). Interestingly, at high levels of ZAP-S or ZAP-L, the 36 introduced CpGs in env and the 53 CpGs in pol inhibited HIV-1 infectious virus production to similar levels (Fig. 7A). Both HIV-1-IRES-GFP and HIV-1-IRES-Renilla were potently inhibited by overexpressed ZAP, indicating that when ZAP abundance is high enough, the CpGs introduced in these viruses can be targeted (Fig. 7B). Therefore, the inhibition observed for HIV-1-IRES-GFP and HIV-1-IRES-Renilla at high ZAP levels indicates that CpGs in contexts that are not targeted by the endogenous ZAP levels in HeLa cells (Fig. 4) can be targeted if ZAP abundance is substantially increased (Fig. 7B).

**CpG dinucleotides introduced into the 5' end of gag inhibit HIV-1 replication in a ZAP-independent manner.**

We had previously introduced CpG dinucleotides into nt 22-378 of gag in two different contexts (Fig. 8A, Table 1) and found that they inhibited viral replication with different phenotypes on Gag expression (15). For HIV-1_{gag22-378}CM, the modified sequence was derived from a codon-optimized Gag-Pol construct and introduced 109 synonymous nucleotide changes and 26 CpGs (Table 1) (15, 46). In the context of a single cycle infectivity assay, HIV-1_{gag22-378}CM Gag expression and infectious virus production were decreased to the limit of detection (Fig. 8B). This correlated with a large decrease in genomic RNA abundance in the cell lysate and media (Fig. 8C-D). The CpG dinucleotides were necessary to inhibit the virus because when they were removed (while leaving the 79 mutations that did not introduce a CpG
dinucleotide) to create HIV-1\textsubscript{gag22-378}CM-no-CpG (Fig. 8A, Table 1), gRNA abundance, Gag expression and infectious virion production were substantially increased (Fig. 8B). However, when the 26 CpG dinucleotides were introduced without the additional mutations in the codon-optimized Gag sequence to create HIV-1\textsubscript{gag22-378}CpG (Fig. 8A, Table 1), infectious virus production was decreased by >95% even though there was no substantial decrease in Gag expression (Fig. 8B).

Deletion of nt 22-378 in gag does not substantially decrease infectious virus production (15, 47), indicating that there are no essential cis-acting elements in this region. However, altering the RNA sequence could modulate the local RNA structure in ways that a large deletion does not and the 5’ region of gag has been shown to indirectly regulate gRNA packaging by regulating the structure of the 5’ UTR (48, 49).

To determine if the CpGs in HIV-1\textsubscript{gag22-378}CpG specifically decrease infectious virus production or whether a cis-acting regulatory element in this region had been mutated, we changed the codons previously mutated to introduce CpGs to a different codon (DC) that was not the wild type HIV-1 codon and did not introduce a CpG where possible to produce pHIV-1\textsubscript{gag22-378}DC (Fig. 8A, Table 1). HIV-1\textsubscript{gag22-378}DC produced similar levels of infectious virus as wild type HIV-1 (Fig. 8B). We also measured gRNA abundance in the cell lysate and media. HIV-1\textsubscript{gag22-378}CpG had a ~60% and ~80% decrease in gRNA in the lysate and media, respectively, while HIV-1\textsubscript{gag22-378}DC had similar levels of gRNA as wild type HIV-1 (Fig. 8C-D). This indicates that the introduced CpGs are necessary for the reduction in infectious virus production and not a result of mutating essential cis-acting elements in this region that modulate gRNA packaging or other steps of HIV-1 replication. In sum, while HIV-1\textsubscript{gag22-378}CM and HIV-1\textsubscript{gag22-378}CpG have the same 26 introduced CpGs, there is a
much larger decrease in Gag expression and intracellular gRNA abundance for HIV-1\textsubscript{gag22-378CM} and this is due to the introduced CpGs. This suggests that the sequence surrounding the CpG dinucleotides modulates their inhibitory effect.

To test the effect of the sequence proximal to the CpG, we changed all of the mutations in HIV-1\textsubscript{gag22-378CM} that were not within 5 nucleotides of an introduced CpG back to the wild type HIV-1 sequence to produce HIV-1\textsubscript{gag22-378CM-5nt-CpG} (Fig. 8A, Table 1). Interestingly, HIV-1\textsubscript{gag22-378CM-5nt-CpG} Gag expression and infectious virus production were very similar to HIV-1\textsubscript{gag22-378CM} (Fig. 8E). This indicates that the sequence immediately surrounding the CpG dinucleotides influences the degree to which they inhibit HIV-1 gene expression. We also analyzed intracellular Env abundance and found that Env expression was reduced to undetectable levels for HIV-1\textsubscript{gag22-378CM} and HIV-1\textsubscript{gag22-378CM-5nt-CpG} (Fig. 8E). Env expression was also decreased for HIV-1\textsubscript{gag22-378CpG}, which likely accounts for the decreased infectious virus production in addition to the reduced gRNA levels present in virions (Fig. 8D).

The decrease in Env expression was unexpected because the region in the HIV-1 genome containing the introduced CpGs in these viruses is only present in the unspliced transcript encoding Gag and Gag-Pol and not in the singly spliced \textit{env} mRNAs. Overall, the local sequence context of the CpGs in \textit{gag} nt 22-378 determines the magnitude of inhibition for both Gag and Env expression.

We then analyzed whether ZAP was necessary for the CpGs in HIV-1\textsubscript{gag22-378CM} and HIV-1\textsubscript{gag22-378CpG} to inhibit Gag expression and infectious virus production. In contrast to HIV-1\textsubscript{env86-561CpG}, ZAP depletion did not increase infectious virus production for HIV-1\textsubscript{gag22-378CM} and HIV-1\textsubscript{gag22-378CpG} (Fig. 9A). This shows that the
CpGs introduced into the 5′ region of gag inhibit HIV-1 gene expression and infectious virus production through a ZAP-independent mechanism. To determine if introducing a larger number of CpGs into gag could make it ZAP-sensitive, we cloned pHIV-1\_gag22-1188CpG which contains 62 CpG dinucleotides distributed across ~1100 nt (Fig. 9B, Table 1). In addition, we produced pHIV-1\_gag22-654CpG and HIV-1\_gag658-1188CpG that contain 30 and 32 CpGs, respectively. Introducing CpGs into nt 22-654 of gag led to a large decrease in Gag expression, Env expression and infectious virus production (Fig. 9C). Interestingly, Gag expression for HIV-1\_gag22-1188CpG and HIV-1\_gag22-654CpG was increased in the ZAP knockout cells, though this did not affect Env expression or infectious virus production. In contrast, introduction of 32 CpGs into nt 658-1188 of gag had no effect on Gag expression, Env expression or infectious virus production (Fig. 9C). Therefore, we introduced 60 CpGs into this 3′ region of gag (HIV-1\_gag694-1206CpG, Fig. 9B, Table 1) and analyzed the effect on infectious virus production in control and ZAP CRISPR cells (Fig. 9D). The 60 CpGs in the 3′ region of gag inhibited HIV-1 infectious virus production about 2-fold, which is similar to the effect that 53 CpGs had in pol and less than the 36 CpGs in env.

Overall, in some contexts, CpGs introduced in gag allow ZAP to inhibit Gag expression but the number of CpGs does not correlate with the magnitude of the inhibitory effect. Furthermore, the CpGs introduced in the 5′ region of gag inhibit infectious virus production in a ZAP-independent manner.

CpG dinucleotides introduced into gag can inhibit HIV-1 gene expression by modulating pre-mRNA splicing.

The HIV-1 genomic RNA undergoes extensive alternative splicing to mediate expression of all of the viral genes and synonymous mutations in gag have
previously been shown to disrupt HIV-1 splicing (36, 50). Since the CpGs in the 5’
region of gag reduced Env expression in all sequence contexts tested (Fig. 8E and
9C), we speculated that they could affect splicing. Therefore, we analyzed RNA
abundance when progressively longer regions of codon-optimized gag sequence
were added to the virus, which introduced 11, 18 or 26 CpGs (HIV-1_{gag22-165}CM, HIV-
1_{gag22-261}CM and HIV-1_{gag22-376}CM, Fig. 10A, Table 1). We have previously analyzed
these viruses and shown that they are deficient for genomic RNA abundance and
Gag expression (15). A comparison of the total RNA and the genomic RNA indicated
that the abundance of both is decreased by the synonymous mutations (Fig. 10B). To
determine whether the decrease in viral RNA abundance is due to altered splicing,
we used RNA-seq to determine the splice sites that were used for each virus (Fig.
10C, Table 2, Data set S3). This analysis showed that a pre-existing cryptic splice
donor (CD1) was activated. Importantly, this donor is outside of the region in which
the CpGs were introduced (Fig. 10D). The frequency of use of the cryptic splice
donor increased with the length of the codon-optimized gag sequence and coincided
with a decrease in the utilization of the canonical splice donor 1 (SD1). Activation of
this cryptic splice donor increased the length of the first exon incorporated into all of
the spliced viral RNAs to include the gag sequence prior to CD1. This leads to the
incorporation of the Gag initiation codon in every transcript upstream of the canonical
start codon for the encoded protein (Fig. 10C and 10D), which is predicted to result in
inefficient translation of all HIV-1 proteins encoded by a singly or fully spliced mRNA,
including Tat and Rev. This could account for the decrease in total RNA levels,
genomic RNA levels, Gag and Env expression that we observed for HIV-1_{gag22-378}CM
(Fig. 8 and 10 and (15)).
DISCUSSION

There is selection against CpG dinucleotides in many vertebrate RNA viruses and introducing them into viral genomes may allow novel vaccines to be developed (1-5, 9, 10). However, to attenuate viral replication, it is unclear how many CpGs are required, if there is an optimal location to insert CpGs and whether all CpGs inhibit via ZAP. Due to the profound suppression in CpG abundance in HIV-1, we have used it as a model system to analyze how CpG dinucleotides inhibit viral replication.

Our results show that CpGs can inhibit HIV-1 replication through at least two independent, but not mutually exclusive, mechanisms. First, they can recruit ZAP and target the viral RNA for degradation. Second, they can inhibit replication by altering pre-mRNA splicing. In addition, CpGs could silence HIV-1 transcription through DNA methylation. The multiple mechanisms by which CpGs inhibit HIV-1 infectious virus production may account for why they are strongly suppressed in this virus, even compared to other RNA viruses, and may also explain why small changes in the number of CpGs in env are linked to disease progression (18). Introducing CpGs into env nt 86-561 potently inhibits genomic RNA abundance, Env expression, Gag expression, and infectious virus production in a ZAP-dependent manner. However, ZAP depletion does not fully rescue infectious virus production when CpGs are introduced in several other regions of the HIV-1 genome, highlighting the ZAP-independent effects of CpGs as well as the sensitivity of the 5' region of env for CpGs that mediate ZAP antiviral activity.

We have characterized the ZAP-independent effect for CpGs introduced in the 5' end of gag. These CpGs have a dramatic effect on genomic RNA levels, Gag expression
and Env expression by promoting the use of a cryptic 5’ splice donor at the expense of SD1. Interestingly, the magnitude of this effect is modulated by the sequence identity immediately surrounding the CpGs. It should be noted that these experiments were done in the context of transiently transfected proviral constructs in HeLa cells, but previous studies have shown that the splicing pattern for HIV-1 is similar in transfected cells and infected T cells (51, 52). The introduced CpGs do not directly enhance the strength of this splice site because they are upstream of the splice donor and do not affect the sequence itself. A previous report has identified that introducing synonymous mutations to the 5’ end of gag promotes splicing at this cryptic donor, but the role of CpGs for this was not characterized (50). We have found that CpGs introduced in this region have multiple effects on viral replication including decreases in genomic RNA stability, Gag expression, virion production and infectivity per genome (15). All of these phenotypes are likely due to the decreased use of SD1 and the corresponding increase in splicing from the cryptic splice donor in gag. The CpGs introduced into the 5’ end of gag could inhibit a pre-existing exonic splicing silencer (ESS) or introduce an exonic splicing enhancer (ESE). We favor the hypothesis that the CpGs have introduced an ESE because the synonymous mutations promote splicing at the cryptic splice donor in a length-dependent manner and the sequence within five nucleotides surrounding the CpG modulates the magnitude of the decrease in Env and Gag expression. However, further experiments will be required to characterize how CpGs modulate splicing. There are approximately 1500 RNA binding proteins in the human genome, most of which do not have a well-characterized recognition sequence, though several have been reported to bind sequences that contain CpGs (53, 54). Therefore, an unknown number of RNA binding proteins bind CpGs and we do not yet know which protein
regulates HIV-1 splicing in a CpG-dependent manner. In addition, introducing CpGs into the HIV-1 genome may affect its local or long-range RNA structures, post-transcriptional modifications such as cytosine methylation or other aspects of RNA biology (55-57).

Surprisingly, the magnitude of ZAP-mediated inhibition did not correlate with the number of CpGs introduced into the viral genome and some regions of the genome can tolerate substantial numbers of CpGs. 36 CpGs inserted into the 5’ region of env had the greatest ZAP-dependent inhibitory phenotype. This corresponds with the observations by Takata et al., who first identified this region in a panel of viruses with large numbers of synonymous mutations in different regions of the HIV-1 genome (16, 50). The introduction of 53 CpGs into pol or 60 CpGs into the 3’ end of gag had only a small ZAP-dependent inhibition on infectious virus production. KHNYN overexpression also had a smaller inhibition of infectious virus production when 53 CpGs were introduced in pol than when 36 CpGs were added to env. This supports our hypothesis that KHNYN antiviral activity is controlled by ZAP’s ability to target the viral RNA (31). Only when ZAP levels were very high due to overexpression from a cDNA plasmid did the 53 CpGs in pol mediate a similar level of repression as the 36 CpGs introduced into env nt 86-561. This highlights that this region in env is very sensitive to the endogenous levels of ZAP in HeLa cells and that the position or local context of the CpG is important for ZAP to inhibit the virus. Interestingly, the weak inhibition of infectious virus production by the 53 CpGs in pol mediated by endogenous ZAP levels could be substantially enhanced by IFN-1 treatment. Therefore, part of the anti-HIV activity mediated by IFN-1 (58) may be to promote ZAP targeting CpGs in contexts that it normally does so inefficiently. Since there is only a
small increase in ZAP-S abundance upon IFN-I treatment, it raises the question of whether this induction of ZAP is sufficient to explain the phenotype or whether increased abundance or activity of ZAP cofactors such TRIM25 or KHNYN may contribute (29-31, 59).

Because HV-1 containing reporter genes are commonly used research tools, we investigated whether the large numbers of CpGs in the EMCV IRES, GFP or Renilla luciferase could sensitize these viruses to ZAP. Adding 96 CpGs to the 3’ end of the genome in the context of IRES-GFP or IRES-Renilla luciferase did not sensitize the virus to the endogenous levels of ZAP in HeLa or HEK293T cells. Similarly, HIV-1 with GFP in the place of nef has previously been shown to not be targeted by endogenous levels of ZAP in HeLa or MT4 cells (16). While HIV-1-IRES-GFP and HIV-1-IRES-Renilla were not inhibited by ZAP after IFN-I treatment, they were inhibited when high levels of ZAP were present due to overexpression from a cDNA plasmid. Therefore, ZAP abundance can determine whether CpG-containing viral genomes are targeted, though it is unclear whether the ZAP levels produced from plasmid-based overexpression can be achieved in a relevant in vivo context. This suggests that these reporter viruses are useful tools that may not be affected by ZAP under many experimental conditions.

An important area of future research is to determine why CpGs in some contexts or regions are efficiently targeted by ZAP and others are not. To date, the primary evidence that ZAP directly binds CpGs is from PAR-CLIP experiments (16). The advantage of this technique is that it captures ZAP binding to CpGs in a living cell. However, other cellular factors present could modulate ZAP’s binding specificity.
Several groups have shown that the ZAP cofactor TRIM25 can bind cellular and viral RNA and it has been reported to regulate ZAP binding to Sindbis virus RNA (29, 30, 60-65). Therefore, TRIM25 or other ZAP cofactors could bind specific motifs in viral RNA to determine the sensitivity of this RNA to ZAP-mediated antiviral activity. In addition, it is not known how many ZAP molecules are required to bind RNA to mediate antiviral activity or if they have to be clustered in a specific way. While structural and mutagenesis studies of the RNA binding domain in ZAP have shown that it is a dimer that may have two RNA-binding cavities within a large RNA-binding cleft, how it binds CpG dinucleotides remains unknown (66, 67). To fully understand how specific CpGs mediate ZAP-dependent antiviral activity, it will be essential to understand how ZAP binds CpGs in specific RNA contexts and structures and the role its cofactors play in modulating its RNA binding activity.

It will be interesting to compare how CpGs inhibit HIV-1 to how they inhibit other RNA viruses and if they do so by targeting viral RNA for degradation, inhibiting its translation or through other mechanisms. CpGs directly or indirectly introduced into coding and non-coding regions of picornaviruses have shown that they can potently attenuate viral replication and create strains that protect animals from challenge with the wild type virus (6-8, 10, 68). ZAP has recently been shown to be necessary for introduced CpGs to inhibit the picornavirus echovirus 7 (32). CpGs have also been shown to attenuate influenza A virus and protect animals from a lethal challenge by the wild type virus (9). However, the molecular mechanism of attenuation remains unclear and the CpGs could inhibit via ZAP, altered RNA splicing similar to what we have observed in HIV-1 or other mechanisms. This highlights both a challenge and opportunity for introducing CpG dinucleotides to create live attenuated vaccines. The
multiple mechanisms of action as well as the position- and context-dependence of 
CpG-mediated viral inhibition pose a challenge to determine the engineering 
principles for attenuating viruses with a predicted magnitude and mechanism. The 
opportunity is that CpGs can be used to attenuate viruses through multiple and 
potentially additive or synergistic mechanisms, which may enhance the utility of this 
approach.

Materials and methods

Sequence analysis of viral genomes
The “analyze base composition” tool in MacVector was used to calculate the CpG 
and UpA observed/expected for the viral sequences. The observed/expected was 
calculated using the following formula: number of dinucleotide occurrences / 
(frequency of the first nucleotide * frequency of second nucleotide) where frequency 
of the base is number of occurrences of the base / total number of bases in 
sequence. For the retroviral sequences, the sequence representing the packaged 
genome was used, i.e. the sequence encompassing the 5’ repeat (R) to the 3’ R.

Synonymous site conservation analysis
Synonymous site conservation was calculated with synplot2 (37) for codons in a gag-
pol-vif-vpr-5'tat-vpu-env-nef in-frame fusion. The 16-nt non-coding sequence 
between tat and vpu was deleted, the frameshift heptanucleotide was changed from 
UUUUUA to UUUUUUAA to fuse gag and pol and single-nt insertions (N) or 
deletions were applied as necessary to fuse ORFs in-frame. Sequences were 
translated to amino acid sequences that were aligned with CLUSTAL W (69). The 
amino acid alignment was then converted back to a nucleotide alignment with
tranalign from the EMBOSS package (70). The alignment was then mapped to the coordinates of a specific reference sequence (viz. EU541617) by removing alignment positions that contained a gap character in the reference sequence. After calculating the synonymous site conservation statistics with synplot2, they were mapped to EU541617 genome coordinates (i.e. by reinserting non-coding regions) for display (Figure 2). The following GenBank sequences were used in the alignment:

- AB023804, AF321523, EU861977, JN944905, JX236669, M38429, AB287379
- AY805330, FJ670516, JN944907, JX236670, AB485648, FJ771006, JN944941
- JX236671, AB485649, FJ771007, JN944942, JX236672, AF004394, FJ771008
- JN944943, JX236673, U26942, AF004885, FJ771009, JN944944, JX236676
- U34603, AF005494, FJ771010, JN944945, JX236677, U34604, AF005496
- GQ290462, JN944946, JX236678, U39362, AF077336, GU237072, JN944947
- JX236679, U63632, AF082394, DQ676872, GU362882, JN944948, K03454
- U88824, AF082395, DQ979025, JN106043, JX236668, M27323, AF286237
- EU541617, KC156212, KC156213, KC156214, KC156215, KC156218, and KC156221. We also used the WARO, MCST, RHGA, STCO1 and CH457 sequences that have previously been described (71, 72). The z-values for 9-, 15- and 25-codon windows are: 9-codon window: $z \geq 3.603808$; 15-codon window: $z \geq 3.468872$; 25-codon window: $z \geq 3.329161$. These equate to $p = 0.05$ with a correction for multiple testing based on the number of independent (i.e. non-overlapping) windows of length $x$ codons in the 2870-codon alignment, i.e. $0.05/(2870/x)$. The scores correspond to the whole $x$-codon window centered on a given codon.

**Plasmids**
All primers were ordered from Eurofins and the PCR reactions were performed with Phusion High Fidelity polymerase (New England Biolabs). Diagnostic restriction enzyme digestion and DNA sequencing (Eurofins Genomics) was used to ensure the correct identity of modified sequences inserted into plasmids. The pHIV-1<sub>NL4-3</sub> constructs used in this study contain the provirus sequence from pHIV-1<sub>NL4-3</sub> (73) cloned into pGL4.10 (Promega) (15). The sequences for HIV-1 gag<sub>22</sub>-378DC, pHIV-1 gag<sub>22</sub>-654CpG, pHIV-1 gag<sub>654</sub>-1188CpG, pHIV-1 gag<sub>694</sub>-1204CpG, pHIV<sub>env</sub>611-1014CpG and pHIV-1 pol<sub>795</sub>-1386CpG were synthesized by Life Technologies and cloned into pHIV-1<sub>NL4-3</sub>. CpG dinucleotides were introduced through synonymous mutations without altering cryptic splice sites predicted by the MaxEnt tool in Human Splicing Finder (74, 75) in pHIV-1 gag<sub>22</sub>-654CpG, pHIV-1 gag<sub>654</sub>-1188CpG, pHIV-1 gag<sub>694</sub>-1204CpG, pHIV<sub>env</sub>611-1014CpG and pHIV-1 pol<sub>795</sub>-1386CpG. To generate pHIV-1 Gag-Venus, the SphI and XmaI fragment of pGag-Venus (43) was amplified via PCR and sub-cloned into pHIV-1<sub>NL4-3</sub> in pGL4.10. pHIV-1 Gag-STOP-Venus was generated by PCR amplification of the SphI and XmaI sites in pGag-Venus using a primer that introduced the 3 stop codons, which was then sub-cloned into pHIV-1Gag-Venus. HA-Renilla, HA-ZAP-S and HA-ZAP-L sequences were amplified by PCR and cloned into pcDNA3.1 (+) (Invitrogen). pHIV-1 gag<sub>22</sub>-165CM, pHIV-1 gag<sub>22</sub>-261CM, pHIV-1 gag<sub>22</sub>-378CM, HIV-1 gag<sub>22</sub>-378CM-no-CpG, pHIV-1 gag<sub>22</sub>-378CpG, pHIV-1 IRES-GFP, pHIV-1 IRES-Renilla, pHIV-1 env<sub>86-561</sub>CpG, pKHNYN-1, pGFP-FLAG, pGFP and pVSV-G have been previously described (15, 31, 41, 42, 76, 77). Of note, pHIV-1 gag<sub>22</sub>-165CM, pHIV-1 gag<sub>22</sub>-261CM, pHIV-1 gag<sub>22</sub>-378CM were previously called pHIV-1 CM22-165, pHIV-1 CM22-261 and pHIV-1 CM22-378, respectively (15). pHIV-1 gag<sub>22</sub>-378CM-no-CpG was previously called pHIV-1 CM22-378<sub>lowCpG</sub> and
pHIV-1gag22-378CpG was called pHIV-1 CpG22-378 (15). pHIV-1env86-561CpG was previously called pHIV-1envCpG86-561 (31).

**Cell culture and cell lines**

HeLa, TZM-bl and HEK293T cells were cultured in Dulbecco’s Modified Eagle Medium (Gibco) supplemented with 10% FBS and 1% penicillin-streptomycin. All cell lines were grown at 37°C in a humidified atmosphere with 5% CO₂. For the production of ZAP and KHNYN knock out cell lines by CRISPR-Cas9 ZAP and KHNYN targeting guide sequences were inserted into a lentivirus-based CRISPR plasmid (LentiCRISPRv2-puro) from Addgene (52961) (78). The CRISPR guide sequences are: Luciferase-G1 (Control): 5'-CTT TAC CGA CGC ACA TAT CG-3', ZAP-ex4: 5'-TCTGGTAGAAGTTATATCTG-3', ZAP-ex6: 5'-ACT TCC ATC TGC CTT ACC GG-3', KHNYN-ex3: 5'-GGG GGT GAG CGT CCT TCC GA-3. LentiCRISPR vectors encoding guide RNAs targeting Luciferase or ZAP were produced in HEK293T cells seeded in a 6-well plate using 10 µl PEI with 0.5µg pVSV-G (76), 1.0 µg pCMVΔR8.91 (79), and 1.0 µg LentiCRISPRv2 (78). Virus containing supernatant was harvested 48-hr after transfection, rendered cell-free via filtration through 0.45 µM filters (Millipore) and used to transduce HeLa or HEK293T cells followed by selection in 1µg/ml puromycin (Sigma-Aldrich). ZAP CRISPR HeLa cell clones were generated by limiting dilution. The ZAP-ex6 and KHNYN-ex3 CRISPR clones have previously been described and were called ZAP-G1 CRISPR clone and KHNYN-G1 CRISPR Clone B (31). Loss of ZAP protein expression was confirmed by western blotting.

**Transfections**
HeLa and HEK293T cells were seeded in a 6 well plate and transfected at a confluence of 70%. HeLa cells were transfected using TransIT-LT1 (Mirus) according to the manufacturer’s instructions at the ratio of 3 μL TransIT-LT1 to 1μg DNA. HEK293T cells were transfected according to the manufacturer’s instructions using PEI (1mg/mL) (Sigma-Aldrich) at the ratio of 4 μL PEI to 1 μg DNA. For each transfection, 0.5μg pHIV-1 and 0.5μg pGFP, pGFP-FLAG, pKHNYN-1, pVSVG, pHA-ZAP-S, pHA-ZAP-L or pHA-Renilla was used for a total of 1μg DNA. The transfection medium was replaced with fresh media 6 hours (HEK293T) or 24 hours (HeLa) post-transfection. The cells were lysed 48 hours post-transfection and the media was recovered. In experiments performed with type-I interferon, 1000 U/mL of IFN-I (PBL 1107 Interferon Source) was added to the cells upon media change 6-hr post-transfection. The media was filtered through a 0.45 µM filter and virions were pelleted for 2-hr at 20,000 x g through a 20% sucrose cushion in phosphate-buffered saline (PBS) solution.

**TZM-bl infectivity assay**

Supernatant was recovered 48 h post-transfection and filtered as previously described. TZM-bl cells were seeded at 70% confluency in 24-well plates and infected by overnight incubation with filtered virus stocks. 48-hours post-infection, the cells were lysed and the amount of infectious virus production was measured by induction of β-galactosidase using the Galacto-Star™ System following manufacturer’s instructions (Applied Biosystems). β-galactosidase activity was quantified as relative light units per second using a PerkinElmner Luminometer.

**Analysis of protein expression by immunoblotting**
Approximately 48-hours post-transfection, HeLa or HEK293T cells were lysed in radioimmunoprecipitation assay (RIPA) buffer (10mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.1% SDS, 1% Triton X-100, 1% sodium deoxycholate). The media was filtered through a 0.45 μm filter and the virions were pelleted through a 20% sucrose cushion in phosphate-buffered saline (PBS) solution for 2 hours at 20,000 x g. The pellet was resuspended in 2x loading buffer (60 mM Tris-HCl (pH 6.8), 10% β-mercaptoethanol, 10% glycerol, 2% sodium dodecyl sulfate (SDS), 0.1% bromophenol blue). Cell lysates and virions were resolved by SDS-polyacrylamide gel electrophoresis and transferred to a nitrocellulose membrane. Antibodies used in the study were 1:50 anti-HIV-1 p24\textsuperscript{Gag} (Mouse, NIH AIDS Reagent Program Cat #1513) (80), 1:3000 anti-HIV-1 gp160/120 Rabbit (ADP421; Centralized Facility for AIDS Reagents (CFAR), 1:1000 anti-Hsp90 (sc7947: Santa Cruz Biotechnology), 1:5000 β-actin (ac-15: Sigma), 1:5000 anti-ZAP (Abcam, ab154680), 1:2500 anti-HA (Biolegend, 901514), 1:2500 anti-DYKDDDDK (Rabbit) (Cell Signalling, 14793), 1:1000 anti-HIV-1 Nef (Mouse, NIH AIDS Reagent Program Cat #1539) (81), 1:10,000 Dylight\textsuperscript{TM} 800-conjugated anti-mouse/rabbit secondary antibodies (Cell Signalling Technology, 5151S and 5257S), 1:5000 anti-mouse IRDye 680RD (LI-COR, 926-68070), 1:5000 anti-rabbit HRP (Cell Signalling Technology, 7074), and 1:5000 anti-mouse HRP (Cell Signalling Technology, 7076). Bound primary antibodies were detected via Li-CoR infrared imaging (LI-COR UK LTD) or using the Amersham ECL Prime Western Blotting Detection reagent (GE Lifesciences) for HRP-linked antibodies using an ImageQuant (LAS4000 Mini).

Quantitative RT-PCR

HeLa cells were transfected at a confluency of 70% in a 6 well plate and after 48h were washed with 1x PBS and lysed. The RNA was extracted using the RNeasy kit.
(QIAGEN) following the manufacturer’s instructions. The supernatant was also collected and treated for 3h at 37°C with RQ1 DNase (Invitrogen) to decrease plasmid DNA contamination. cDNA was synthesized using the High Capacity cDNA archive kit (Applied Biosystems) and using 1μg of RNA from virions was isolated using QIAamp viral RNA mini kit following the manufacturer’s instructions. Because carrier RNA is added to the lysis buffer, the total RNA isolated was quantified using a Qubit 3.0 fluorometer (ThermoFisher) and normalized so that 20 ng of RNA from each sample was reverse transcribed using the High Capacity cDNA archive kit (Applied Biosystems). qPCR reactions were performed in triplicate with Taqman Universal PCR mix using the Applied Biosystems 7500 real-time PCR system. HIV-1NL4-3 gRNA primers were GGCCAGGGAATTTTCTTCAGA/TTGTCTCTTCCCACACCTGA (forward/reverse) and the probe was FAM-ACACAACAGACGGGCACACTA-TAMRA. HIV-1NL4-3 total RNA primers were TAACTAGGGAACCCACTGC/ GCTAGAGATTTTCCACACTG (forward/reverse) and the probe was FAM-ACACAACAGACGGGCACACTA-TAMRA. The absolute number of copies was quantified using the slope of the standard curve and at a qPCR efficiency between 95%-105%.

Analysis of HIV-1 splicing

HeLa cells were transfected at a confluency of 70% in a 6-well plate and after ~40 hours were washed with 1x PBS and lysed. The RNA was extracted using the RNeasy kit (QIAGEN) following the manufacturer’s instructions. The RNA samples were quantified using a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and RNA integrity was checked with RNA Screen Tape on the Agilent 2200 TapeStation (Agilent Technologies, Palo Alto, CA, USA). The multiplexed RNA
sequencing library preparation was prepared using an Illumina TruSeq Stranded mRNA library Prep kit following the manufacturer’s protocol. Sequencing libraries were validated using DNA Analysis Screen Tape on the Agilent 2200 TapeStation and quantified by using Qubit 2.0 Fluorometer as well as by quantitative PCR (Applied Biosystems, Carlsbad, CA, USA). The sequencing libraries were multiplexed and clustered on two flowcell lanes. After clustering, the flowcell was loaded on the Illumina HiSeq instrument according to manufacturer’s instructions. The samples were sequenced using a 2x150 Pair-End (PE) High Output configuration by GENEWIZ. Image analysis and base calling were conducted by the HiSeq Control Software (HCS) on the HiSeq instrument. The raw sequence data (.bcl files) generated from Illumina HiSeq was converted into fastq files and de-multiplexed using the Illumina bcl2fastq program version 2.17. Adapter trimming was performed using BBduk (https://jgi.doe.gov/data-andtools/bbtools/), and read pairs were merged with BBmerge (https://jgi.doe.gov/data-and-tools/bbtools/) in order to increase base call quality and generate long, whole fragment reads. Reads were then aligned to the human genome (hg38) and the HIV\textsubscript{NL4-3} genomic RNA sequence simultaneously using Hisat2 (82). HIV-mapping junction spanning reads were isolated using regtools (https://github.com/griffithlab/regtools) to allow per junction read counting. To visualize junctions of interest, data from replicates were first merged using the Picard (http://broadinstitute.github.io/picard) MergeSamFiles function, followed by generation of sashimi plots using Gviz (83). The percentage HIV-1 junction spanning reads was calculated by dividing the number of reads for each junction by the total number of junction spanning reads in the library.

**Statistical analysis**
Statistical significance was determined using unpaired two-tailed t tests calculated using Microsoft Excel software. Data are represented as mean ± standard deviation. Significance was ascribed to p values < 0.05.

Data availability
The RNA-seq data are available in the NCBI Gene Expression Omnibus (GEO) under accession number GSE134159. The GenBank accession numbers for the HIV-1 genomic RNA nucleotide sequences are as follows: HIV-1 (strain NL4-3), MN685337; HIV-1gag22-165CM, MN685338; HIV-1gag22-261CM, MN685339; HIV-1gag22-375CM, MN685340; HIV-1gag22-375CM-no-CpG, MN685341; HIV-1gag22-375CpG, MN685342; HIV-1gag22-375DC, MN685343; HIV-1gag22-375CM-5nt-CpG, MN685344; HIV-1gag22-651CpG, MN685345; HIV-1gag60-1185CpG, MN685346; HIV-1gag22-1185CpG, MN685347; HIV-1gag694-1206CpG, MN685348; HIV-1pol795-1386CpG, MN685349; HIV-1env86-561CpG, MN685350; HIV-1env611-1014CpG, MN685351; HIV-1env86-1014CpG, MN685352.

Acknowledgements
We thank the members of the Swanson and Neil labs for helpful discussions. The following reagents were obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: TZM-bl from Dr. John C. Kappes, Dr. Xiaoyun Wu and Tranzyme Inc; HIV-1 p24 Hybridoma (183-H12-5C) (Cat# 1513) from Dr. Bruce Chesebro; Anti-HIV-1 NefMonoclonal (6.2) from Dr. Kai Krohn and Dr. Vladimir Ovod. The Antiserum to HIV-1 gp120 #20 (ARP421) was obtained from the NIBSC Centre for AIDS Reagents. We thank Akira Ono and Harmit Malik for generously providing
reagents. We also thank Robert Gifford for providing advice and sequence information on the retroviral phylogeny.

These studies were funded by MRC Discovery Award MC/PC/15068 and a Wellcome Trust Senior Research Fellowship (WT098049AIA) to SJDN, Wellcome Trust Senior Research Fellowship (106207) to AEF, Medical Research Council grant MR/M019756/1 to RS and CMS and Medical Research Council grant MR/S000844/1 to SJDN and CMS. MF is supported by the UK Medical Research Council (MR/R50225X/1) and is a King’s College London member of the MRC Doctoral Training Partnership in Biomedical Sciences. IAA and RHW are supported by the King’s Bioscience Institute and the Guy’s and St Thomas’ Charity Prize PhD Programme in Biomedical and Translational Science. This work was also supported by the Department of Health via a National Institute for Health Research Comprehensive Biomedical Research Centre award to Guy’s and St. Thomas’ NHS Foundation Trust in partnership with King’s College London and King’s College Hospital NHS Foundation Trust.

REFERENCES

1. Greenbaum BD, Levine AJ, Bhanot G, Rabdan R. 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog 4:e1000079.

2. 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.

3. Simmonds P, Xia W, Baillie JK, McKinnon K. 2013. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla--selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. BMC Genomics 14:610.

4. Rima BK, McFerran NV. 1997. Dinucleotide and stop codon frequencies in single-stranded RNA viruses. J Gen Virol 78:2859-2870.
5. Kunec D, Osterrieder N. 2016. Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias. Cell Rep 14:55-67.

6. Atkinson NJ, Witteveldt J, Evans DJ, Simmonds P. 2014. The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. Nucleic Acids Res 42:4527-4545.

7. Burns CC, Campagnoli R, Shaw J, Vincent A, Jorba J, Kew O. 2009. Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons. J Virol 83:9957-9969.

8. Fros JJ, Dietrich I, Alshaikhahmed K, Passchier TC, Evans DJ, Simmonds P. 2017. CpG and UpA dinucleotides in both coding and non-coding regions of echovirus 7 inhibit replication initiation post-entry. Elife 6:e29112.

9. Gaunt E, Wise HM, Zhang H, Lee LN, Atkinson NJ, Nicol MQ, Highton AJ, Klenerman P, Beard PM, Dutia BM, Digard P, Simmonds P. 2016. Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. Elife 5:e12735.

10. Tulloch F, Atkinson NJ, Evans DJ, Ryan MD, Simmonds P. 2014. RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. Elife 3:e04531.

11. Frankel AD, Young JA. 1998. HIV-1: fifteen proteins and an RNA. Annu Rev Biochem 67:1-25.

12. Kypr J, Mrazek J, Reich J. 1989. Nucleotide composition bias and CpG dinucleotide content in the genomes of HIV and HTLV 1/2. Biochim Biophys Acta 1009:280-282.

13. Shpaer EG, Mullins JI. 1990. Selection against CpG dinucleotides in lentiviral genes: a possible role of methylation in regulation of viral expression. Nucleic Acids Res 18:5793-5797.

14. Ohno S, Yomo T. 1990. Various regulatory sequences are deprived of their uniqueness by the universal rule of TA/CG deficiency and TG/CT excess. Proc Natl Acad Sci U S A 87:1218-1222.

15. Antzin-Anduetza I, Mahiet C, Granger LA, Odendall C, Swanson CM. 2017. Increasing the CpG dinucleotide abundance in the HIV-1 genomic RNA inhibits viral replication. Retrovirology 14:49.

16. Takata MA, Goncalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, Bieniasz PD. 2017. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature 550:124-127.

17. Theys K, Feder AF, Gelbart M, Hartl M, Stern A, Pennings PS. 2018. Within-patient mutation frequencies reveal fitness costs of CpG dinucleotides and drastic amino acid changes in HIV. PloS Genet 14:e1007420.

18. Wasson MK, Borkakoti J, Kumar A, Biswas B, Vivekanandan P. 2017. The CpG dinucleotide content of the HIV-1 envelope gene may predict disease progression. Sci Rep 7:8162.

19. Alinejad-Rokny H, Anwar F, Waters SA, Davenport MP, Ebrahim D. 2016. Source of CpG Depletion in the HIV-1 Genome. Mol Biol Evol 33:3205-3212.

20. Gao G, Guo X, Goff SP. 2002. Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein. Science 297:1703-1706.
21. **Bick MJ, Carroll JW, Gao G, Goff SP, Rice CM, MacDonald MR.** 2003. Expression of the zinc-finger antiviral protein inhibits alphavirus replication. *J Virol* 77:11555-11562.

22. **Muller S, Moller P, Bick MJ, Wurr S, Becker S, Gunther S, Kummerer BM.** 2007. Inhibition of filovirus replication by the zinc finger antiviral protein. *J Virol* 81:2391-2400.

23. **Mao R, Nie H, Cai D, Zhang J, Liu H, Yan R, Cuconati A, Block TM, Guo JT, Guo H.** 2013. Inhibition of hepatitis B virus replication by the host zinc finger antiviral protein. *PLoS Pathog* 9:e1003494.

24. **Chiu HP, Chiu H, Yang CF, Lee YL, Chiu FL, Kuo HC, Lin RJ, Lin YL.** 2018. Inhibition of Japanese encephalitis virus infection by the host zinc-finger antiviral protein. *PLoS Pathog* 14:e1007166.

25. **Li M, Yan K, Wei L, Yang J, Xu W.** 2015. Zinc finger antiviral protein inhibits coxsackievirus B3 virus replication and protects against viral myocarditis. *Antiviral Res* 123:50-61.

26. **Guo X, Carroll JW, Macdonald MR, Goff SP, Gao G.** 2004. The zinc finger antiviral protein directly binds to specific viral mRNAs through the CCCH zinc finger motifs. *J Virol* 78:12781-12787.

27. **Guo X, Ma J, Sun J, Gao G.** 2007. The zinc-finger antiviral protein recruits the RNA processing exosome to degrade the target mRNA. *Proc Natl Acad Sci U S A* 104:151-156.

28. **Zhu Y, Wang X, Goff SP, Gao G.** 2012. Translational repression precedes and is required for ZAP-mediated mRNA decay. *EMBO J* 31:4236-4246.

29. **Li MM, Lau Z, Cheung P, Aguilar EG, Schneider WM, Bozzacco L, Molina H, Buehler E, Takaoka A, Rice CM, Felsenfeld DP, MacDonald MR.** 2017. TRIM25 Enhances the Antiviral Action of Zinc-Finger Antiviral Protein (ZAP). *PLoS Pathog* 13:e1006145.

30. **Zheng X, Wang X, Tu F, Wang Q, Fan Z, Gao G.** 2017. TRIM25 Is Required for the Antiviral Activity of Zinc Finger Antiviral Protein. *J Virol* 91:e00088-17.

31. **Ficarelli M, Wilson H, Pedro Galao R, Mazzon M, Antzin-Anduetza I, Marsh M, Neil SJ, Swanson CM.** 2019. KHNYN is essential for the zinc finger antiviral protein (ZAP) to restrict HIV-1 containing clustered CpG dinucleotides. *Elife* 8:e46767.

32. **Odon V, Fros JJ, Goonawardane N, Dietrich I, Ibrahim A, Alshaikhahmed K, Nguyen D, Simmonds P.** 2019. The role of ZAP and OAS3/RNAseL pathways in the attenuation of an RNA virus with elevated frequencies of CpG and UpA dinucleotides. *Nucleic Acids Res* 47:8061-8083.

33. **Brierley I, Dos Ramos FJ.** 2006. Programmed ribosomal frameshifting in HIV-1 and the SARS-CoV. *Virus Res* 119:29-42.

34. **Pollard VW, Malim MH.** 1998. The HIV-1 Rev protein. *Annu Rev Microbiol* 52:491-532.

35. **Le Grice SF.** 2012. Human immunodeficiency virus reverse transcriptase: 25 years of research, drug discovery, and promise. *J Biol Chem* 287:40850-40857.

36. **Sertznig H, Hillebrand F, Erkelenz S, Schaal H, Widera M.** 2018. Behind the scenes of HIV-1 replication: Alternative splicing as the dependency factor on the quiet. *Virology* 516:176-188.

37. **Firth AE.** 2014. Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. *Nucleic Acids Res* 42:12425-12439.
38. Kass SU, Goddard JP, Adams RL. 1993. Inactive chromatin spreads from a focus of methylation. Mol Cell Biol 13:7372-7379.
39. Hsieh CL. 1994. Dependence of transcriptional repression on CpG methylation density. Mol Cell Biol 14:5487-5494.
40. Hsieh CL. 1997. Stability of patch methylation and its impact in regions of transcriptional initiation and elongation. Mol Cell Biol 17:5897-5904.
41. Munch J, Rajan D, Schindler M, Specht A, Rucker E, November FJ, Nerrienet E, Muller-Trutwin MC, Peeters M, Hahn BH, Kirchhoff F. 2007. Nef-mediated enhancement of virion infectivity and stimulation of viral replication are fundamental properties of primate lentiviruses. J Virol 81:13852-13864.
42. Manganaro L, de Castro E, Maestre AM, Olivieri K, Garcia-Sastre A, Fernandez-Sesma A, Simon V. 2015. HIV Vpu Interferes with NF-κB Activity but Not with Interferon Regulatory Factor 3. J Virol 89:9781-9790.
43. Hayakawa S, Shiratori S, Yamato H, Kameyama T, Kitatsui C, Kashigi F, Goto S, Kameoka S, Ohba Y, Miyazaki T, Imamura M, Takaoka A. 2011. ZAPS is a potent stimulator of signaling mediated by the RNA helicase RIG-I during antiviral responses. Nat Immunol 12:37-44.
44. Goodier JL, Pereira GC, Cheung LE, Rose RJ, Kazazian HH, Jr. 2015. The Broad Spectrum Antiviral Protein ZAP Restricts Human Retrotransposition. PLoS Genet 11:e1005252.
45. Mostoslavsky G, Kotton DN, Fabian AJ, Gray JT, Lee JS, Mulligan RC. 2005. Efficiency of transduction of highly purified murine hematopoietic stem cells by lentiviral and oncoretroviral vectors under conditions of minimal in vitro manipulation. Mol Ther 11:932-940.
46. Reil H, Bukovsky AA, Gelderblom HR, Gottlinger HG. 1998. Efficient HIV-1 replication can occur in the absence of the viral matrix protein. EMBO J 17:2699-2708.
47. Liu Y, Nikolaitchik OA, Rahman SA, Chen J, Pathak VK, Hu WS. 2017. HIV-1 Sequence Necessary and Sufficient to Package Non-viral RNAs into HIV-1 Particles. J Mol Biol 429:2542-2555.
48. Kharytonchyk S, Brown JD, Sterliger K, Yasin S, Iyer AS, Collins J, Summers MF, Telesnitsky A. 2018. Influence of gag and RRE Sequences on HIV-1 RNA Packaging Signal Structure and Function. J Mol Biol 430:2066-2079.
49. Takata MA, Soll SJ, Emery A, Blanco-Melo D, Swanstrom R, Bieniasz PD. 2018. Global synonymous mutagenesis identifies cis-acting RNA elements that regulate HIV-1 splicing and replication. PLoS Pathog 14:e1006824.
50. Emery A, Zhou S, Pollom E, Swanstrom R. 2017. Characterizing HIV-1 Splicing by Using Next-Generation Sequencing. J Virol 91.
51. Purcell DF, Martin MA. 1993. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. J Virol 67:6365-6378.
52. Gerstberger S, Hafner M, Tuschi T. 2014. A census of human RNA-binding proteins. Nat Rev Genet 15:829-845.
Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irinima M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecenas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Lei EP, Fraser AG, Blencowe BJ, Morris JD, Hughes TR. 2013. A compendium of RNA-binding motifs for decoding gene regulation. Nature 499:172-177.

McIntyre W, Netzband R, Bonenfant G, Biegel JM, Miller C, Fuchs G, Henderson E, Arra M, Canki M, Fabris D, Pager CT. 2018. Positive-sense RNA viruses reveal the complexity and dynamics of the cellular and viral epitranscriptomes during infection. Nucleic Acids Res 46:5776-5791.

Lavender CA, Gorelick RJ, Weeks KM. 2015. Structure-Based Alignment and Consensus Secondary Structures for Three HIV-Related RNA Genomes. PLoS Comput Biol 11:e1004230.

Courtney DG, Tsai K, Bogerd HP, Kennedy EM, Law BA, Emery A, Swanstrom R, Hooley CL, Cullen BR. 2019. Epitranscriptomic Addition of m(5)C to HIV-1 Transcripts Regulates Viral Gene Expression. Cell Host Microbe 26:217-227.

Doyle T, Goujon C, Malim MH. 2015. HIV-1 and interferons: who's interfering with whom? Nat Rev Microbiol 13:403-413.

Shaw AE, Hughes J, Gu Q, Behdenna A, Singer JB, Dennis T, Orton RJ, Varela M, Gifford RJ, Wilson SJ, Palmarini M. 2017. Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses. PLoS Biol 15:e2004086.

Choudhury NR, Heikel G, Trubitsyna M, Kubik P, Nowak JS, Webb S, Granneman S, Spanos C, Rappasilber J, Castello A, Michlewski G. 2017. RNA-binding activity of TRIM25 is mediated by its PRY/SPRY domain and is required for ubiquitination. BMC Biol 15:105.

Kwon SC, Yi H, Eichelbaum K, Fohr S, Fischer B, You KT, Castello A, Krijgsveld J, Hentze MW, Kim VN. 2013. The RNA-binding protein repertoire of embryonic stem cells. Nat Struct Mol Biol 20:1122-1130.

Manokaran G, Finol E, Wang C, Gunaratne J, Bahl J, Ong EZ, Tan HC, Sessions OM, Ward AM, Gubler DJ, Harris E, Garcia-Blanco MA, Ooi EE. 2015. Dengue subgenomic RNA binds TRIM25 to inhibit interferon expression for epidemiological fitness. Science 350:217-221.

Sanchez JG, Sparrer KJ, Chiang C, Reis RA, Chiang JJ, Zurenski MA, Wan Y, Gack MU, Pornillos O. 2018. TRIM25 Binds RNA to Modulate Cellular Anti-viral Defense. J Mol Biol 430:5280-5293.

Choudhury NR, Nowak JS, Zuo J, Rappasilber J, Spoel SH, Michlewski G. 2014. Trim25 is an RNA-Specific Activator of Lin28a/TuT4-Mediated Uridylation. Cell Rep 9:1265-1272.

Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, Krijgsveld J, Hentze MW. 2012. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell 149:1393-1406.

Law LM, Albin OR, Carroll JW, Jones CT, Rice CM, Macdonald MR. 2010. Identification of a dominant negative inhibitor of human zinc finger antiviral protein reveals a functional endogenous pool and critical homotypic interactions. J Virol 84:4504-4512.
67. Chen S, Xu Y, Zhang K, Wang X, Sun J, Gao G, Liu Y. 2012. Structure of N-terminal domain of ZAP indicates how a zinc-finger protein recognizes complex RNA. Nat Struct Mol Biol 19:430-435.

68. Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. 2008. Virus attenuation by genome-scale changes in codon pair bias. Science 320:1784-1787.

69. Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680.

70. Rice P, Longden I, Bleasby A. 2000. EMBOS: the European Molecular Biology Open Software Suite. Trends Genet 16:276-277.

71. Parrish NF, Gao F, Li H, Giorgi EE, Barbian HJ, Parrish EH, Zajic L, Iyer SS, Decker JM, Kumar A, Hora B, Berg A, Cai F, Hopper J, Denny TN, Ding H, Ochsenbauer C, Kappes JC, Galimidi RP, West AP, Jr., Bjorkman PJ, Wilen CB, O'Brien M, Hahn BH. 2013. Phenotypic properties of transmitted founder HIV-1. Proc Natl Acad Sci U S A 110:6626-6633.

72. Parrish NF, Wilen CB, Banks LB, Iyer SS, Pfaff JM, Salazar-Gonzalez JF, Salazar MG, Decker JM, Parrish EH, Berg A, Hopper J, Hora B, Kumar A, Mahlokozera T, Yuan S, Coleman C, Vermeulen M, Ding H, Ochsenbauer C, Tilton JC, Permra SR, Kappes JC, Betts MR, Busch MP, Gao F, Montefiori D, Haynes BF, Shaw GM, Hahn BH, Doms RW. 2012. Transmitted/founder and chronic subtype C HIV-1 use CD4 and CCR5 receptors with equal efficiency and are not inhibited by blocking the integrin alpha4beta7. PLoS Pathog 8:e1002686.

73. Adachi A, Gendelman HE, Koenig S, Folks T, Willey R, Rabson A, Martin MA. 1986. Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. J Virol 59:284-291.

74. Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11:377-394.

75. Desmet FO, Hamroun D, Lalande M, Collo-Beroud G, Claustres M, Beroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37:e67.

76. Fouchier RA, Meyer BE, Simon JH, Fischer U, Malim MH. 1997. HIV-1 infection of non-dividing cells: evidence that the amino-terminal basic region of the viral matrix protein is important for Gag processing but not for post-entry nuclear import. EMBO J 16:4531-4539.

77. Swanson CM, Sherer NM, Malim MH. 2010. SRp40 and SRp55 promote the translation of unspliced human immunodeficiency virus type 1 RNA. J Virol 84:6748-6759.

78. Sanjana NE, Shalem O, Zhang F. 2014. Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods 11:783-784.

79. Zufferey R, Nagy D, Mandel RJ, Naldini L, Trono D. 1997. Multiply attenuated lentiviral vector achieves efficient gene delivery in vivo. Nat Biotechnol 15:871-875.

80. Chesebro B, Wehrly K, Nishio J, Perryman S. 1992. Macrophage-tropic human immunodeficiency virus isolates from different patients exhibit unusual V3 envelope sequence homogeneity in comparison with T-cell-tropic isolates: definition of critical amino acids involved in cell tropism. J Virol 66:6547-6554.
Ovod V, Lagerstedt A, Ranki A, Gombert FO, Spohn R, Tahtinen M, Jung G, Krohn KJ. 1992. Immunological variation and immunohistochemical localization of HIV-1 Nef demonstrated with monoclonal antibodies. AIDS 6:25-34.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357-360.

Hahne F, Ivanek R. 2016. Visualizing Genomic Data Using Gviz and Bioconductor. Methods Mol Biol 1418:335-351.

FIGURE LEGENDS

Fig. 1. HIV-1 is highly suppressed for CpG abundance and not substantially suppressed for UpA abundance compared to many viruses that infect vertebrate cells. (A) The number of CpGs/kilobase (kb) and the CpG observed/expected ratio were calculated for each virus. The members of the Retroviridae are shown in red and the other viruses are shown in blue. (B) The numbers of CpGs/kb and CpG observed/expected ratio are plotted for the Retroviridae family. Members of the Lentivirus genus are shown in turquoise and the other retroviruses are shown in red. The viruses naturally found in humans, chimpanzees and gorillas are shown in purple. (C) The number of UpAs/kb and the UpA observed/expected ratio were calculated for each virus. The members of the Retroviridae are shown in red and the other viruses are shown in blue. Members of the Flaviridae family are shown in orange. (D) The numbers of UpAs/kb and UpAs observed/expected ratio are plotted for the Retroviridae family. Members of the Lentivirus genus are shown in turquoise and the other retroviruses are shown in red. The viruses naturally found in humans, chimpanzees and gorillas are shown in purple. See Data set S1 for the values for each virus.

Fig. 2. Overlapping functional elements are localized to specific regions of the HIV-1 genome. (A) Schematic representation of the HIV-1 provirus indicating the open reading frames. (B) Analysis of conservation at synonymous sites. The brown
line depicts the ratio of the observed number of substitutions to the number expected under a null model of neutral evolution at synonymous sites for a 25-codon sliding window. (C-E) Red lines show the corresponding p-values for 25-, 15- and 9-codon sliding window analyses. Grey dashed lines indicate approximate p = 0.05 false positive thresholds after correcting for multiple testing in each plot. See Data set S2 for the Z-scores for each codon in the HIV-1 open reading frames in the 9-, 15- or 25-codon sliding windows. The regions with known cis-acting elements are identified: FS = programmed ribosomal frameshift sequence (33), cPPT = central polypurine tract (35), SRS = splicing regulatory sequences (36), RRE = Rev-response element (34), PPT = polypurine tract (35).

**Fig. 3. Introduction of CpGs in env inhibits HIV-1 infectious virus production through ZAP-dependent and -independent mechanisms.** (A) Schematic representation of the HIV-1 env region with the region of synonymously introduced CpGs in HIV-1env66-561CpG, HIV-1env611-1014CpG, and HIV-1env66-1014CpG highlighted. (B) ZAP expression in HeLa Untransduced, Control, ZAP-ex4 and ZAP-ex6 CRISPR cells was detected using immunoblotting. The untransduced cells are the parental HeLa cells not transduced with a CRISPR-Cas9 lentiviral vector. (C-D) HeLa Control, ZAP-ex4 and ZAP-ex6 CRISPR cells were transfected with pHIV-1, pHIV-1env66-561CpG, pHIV-1env611-1014CpG, or pHIV-1env66-1014CpG plus pGFP (C) or pVSV-G (D). Culture supernatants were used to infect TZM-bl reporter cells to measure infectious virus production. Gag expression in the media, as well as Gag, Hsp90, Env, Actin, and Nef expression in the cell lysates was detected using immunoblotting. The bar charts show the average of four independent experiments normalized to wild type HIV-1 in HeLa Control CRISPR cells. (E) Genomic RNA abundance was quantified...
by qRT-PCR in cell lysates. The bar charts show the average of three independent experiments normalized to wild type HIV-1 in HeLa Control CRISPR cells. Error bars represent standard deviation. *p<0.05 as determined by a two-tailed unpaired t-test. The black * compares the virus containing introduced CpGs in the Control CRISPR cells to wild type HIV-1 in the Control CRISPR cells. The red * compares the virus containing introduced CpGs between the ZAP CRISPR cells and the Control CRISPR cells.

Fig. 4: Introduction of 53 CpGs into pol only moderately sensitizes the virus for ZAP- and KHNYN-dependent inhibition of infectious virus production. (A) Schematic representation of the HIV-1 pol with the region of synonymously introduced CpGs in HIV-1pol795-1386CpG highlighted. (B) HeLa Control, ZAP-ex4 and ZAP-ex6 CRISPR cells were transfected with pHIV-1 or pHIV-1pol795-1386CpG plus pGFP. Viral infectivity was measured using TZM-bl reporter cells infected with cell culture supernatants. Gag expression in the media, as well as Gag and Hsp90 expression in the cell lysates was detected by immunoblotting. The bar charts show the average values of four independent experiments normalized to the values obtained for wild type HIV-1 in the HeLa Control CRISPR cells. (C-D) HeLa cells were transfected with 500 ng pHIV-1, pHIV-1EnvCpG86-561 or pHIV-1pol795-1386CpG and 500 ng of pGFP-FLAG or 31.25 ng, 62.5 ng, 125 ng, 250 ng or 500 ng pKHNYN-1-FLAG plus the amount of pGFP-FLAG required to make 500 ng total. (C) Viral infectivity was measured using TZM-bl reporter cells infected with cell culture supernatants. Each point shows the average value of three independent experiments normalized to the value obtained for wild type HIV-1 in HeLa cells. *p<0.05 as determined by a two-tailed unpaired t-test. The black * compares the virus containing
introduced CpGs to wild type HIV-1 with 0 ng of KHNYN. The red * compares the virus containing introduced CpGs between each point of KHNYN overexpression to 0 ng of KHNYN. (D) Gag expression in the media as well as Gag, Hsp90, Env, Actin, and KHNYN-FLAG expression in the cell lysates was detected using immunoblotting. (E) HeLa Control, and KHNYN-ex3 CRISPR cells were transfected with pHIV-1, or pHIV-1\textsubscript{env}86\textsubscript{-}561\textsubscript{CpG}, pHIV-1\textsubscript{env}611\textsubscript{-}1014\textsubscript{CpG}, pHIV-1\textsubscript{pol}795\textsubscript{-}1386\textsubscript{CpG} plus pGFP. Viral infectivity was measured using TZM-bl reporter cells infected with cell culture supernatants. Gag expression in the media, as well as Gag, Hsp90, Env and Actin expression in the cell lysates was detected using immunoblotting. The bar charts show the average values of three independent experiments normalized to the values obtained for wild type HIV-1 in the HeLa Control CRISPR cells. Error bars represent standard deviation. *p<0.05 as determined by a two-tailed unpaired t-test. The black * compares the virus containing introduced CpGs in the Control CRISPR cells to wild type HIV-1 in the Control CRISPR cells. The red * compares the virus containing introduced CpGs between the KHNYN CRISPR cells and the Control CRISPR cells.

Fig. 5: HIV-1-IRES-GFP, HIV-1-IRES-\textit{Renilla} and HIV-1 containing a Gag-Venus fusion protein are not targeted by endogenous levels of ZAP in HeLa cells. (A) Schematic representation of the HIV-1 \textit{nef}/LTR region with the introduced GFP and Renilla reporter genes and HIV-1 Gag-Venus fusion constructs. (B) Control, ZAP-ex4 and ZAP-ex6 CRISPR cells were transfected with pHIV-1, pHIV-1\textsubscript{env}86\textsubscript{-}561\textsubscript{CpG}, pHIV-1-IRES-GFP or pHIV-1-IRES-\textit{Renilla} plus pGFP. Infectious virus production was measured using TZM-bl reporter cells infected with cell culture supernatants. Gag expression in the media, as well as Gag and Hsp90 expression in the cell lysates
was detected using immunoblotting. All bar charts show the average values of three independent experiments normalized to the values obtained for wild type HIV-1 in HeLa Control CRISPR cells. (C) Control and ZAP-ex6 CRISPR cells were transfected with pHIV-1Gag-Venus or pHIV-1Gag-STOP-Venus. Gag expression in the media, as well as Gag and Hsp90 expression in the cell lysates was detected using immunoblotting. The bar charts show the average values of three independent experiments normalized to the virus in Control CRISPR cells. Error bars represent standard deviation. *p<0.05 as determined by a two-tailed unpaired t-test. The black * compares the virus containing introduced CpGs in the Control CRISPR cells to wild type HIV-1 in the Control CRISPR cells. The red * compares the virus containing introduced CpGs between the ZAP CRISPR cells and the Control CRISPR cells.

Fig. 6: Type I interferon inhibits HIV-1 containing introduced CpG dinucleotides using ZAP-dependent and ZAP-independent mechanisms and induces an ~2-fold increase in ZAP-S expression in 293T cells. (A-B) HEK293T Control and ZAP-ex6 CRISPR cells were transfected with pHIV-1, pHIV-1ev561CpG, pHIV-pol795CpG, pHIV-1-IRES-GFP or pHIV-1-IRES-Renilla plus pGFP with or without 1000 U/ml of type I interferon treatment. Infectious virus production was measured using TZM-bl reporter cells infected with cell culture supernatants. Gag expression in the media, as well as Gag and Hsp90 expression in the cell lysates was detected using immunoblotting. The bar charts show the average values of four independent experiments normalized to wild type HIV-1 in HEK293T Control CRISPR cells. The number over the bars represents the fold decrease in relative infectious virus production due to the type I interferon treatment. Error bars represent standard deviation. *p<0.05 as determined by a two-tailed unpaired t-test. The black *
compares the virus containing introduced CpGs in the Control CRISPR cells to wild

Fig. 7: ZAP overexpression inhibits wild type HIV-1 and HIV-1 containing

introduced CpG dinucleotides. (A-B) HeLa cells were transfected with pHIV-1,

pHIV-1<sub>env</sub>66-561CpG, pHIV-1<sub>pol</sub>795-1386CpG, pHIV-1-IRES-GFP or pHIV-1-IRES-Renilla

plus either pH-A-Renilla, pHA-ZAP-S or pHA-ZAP-L. Infectious virus production was

measured using TZM-bl reporter cells infected with cell culture supernatants. Gag

expression in the media, as well as Gag, and Hsp90 expression in the cell lysates

was detected using immunoblotting. The bar charts show the average values of four

independent experiments normalized to wild type HIV-1 plus HA-Renilla. The number

over the bars represents the fold decrease in relative infectious virus production due

to ZAP-S or ZAP-L overexpression. (C) Bar chart and representative western blot for

ZAP expression, showing the average ZAP abundance for 15 independent

transfections of p-HA-ZAP-S or pHA-ZAP-L in HeLa cells. Error bars represent

standard deviation. *p<0.05 as determined by a two-tailed unpaired t-test. The black *

compares the virus containing introduced CpGs in the Control CRISPR cells to wild

type HIV-1 in the Control CRISPR cells. The red * compares the virus with HA-

Renilla overexpression to HA-ZAP overexpression.
Fig. 8. Introduction of CpG dinucleotides into gag nt 22-378 inhibits HIV-1 infectious virus production. (A) Schematic representation of the HIV-1 gag with the introduced synonymous mutations in HIV-1\textsubscript{gag22-378}CM, HIV-1\textsubscript{gag22-378}CpG, HIV-1\textsubscript{gag22-378}CM-no-CpG, HIV-1\textsubscript{gag22-378}DC, and HIV-1\textsubscript{gag22-378}CM-5nt-CpG highlighted. (B-E) HeLa cells were transfected with pHIV-1, pHIV-1\textsubscript{gag22-378}CM, pHIV-1\textsubscript{gag22-378}CM-no-CpG, pHIV-1\textsubscript{gag22-378}CpG or pHIV-1\textsubscript{gag22-378}DC (B-D) or pHIV-1, pHIV-1\textsubscript{gag22-378}CM, pHIV-1\textsubscript{gag22-378}CpG or pHIV-1\textsubscript{gag22-378}CM-5nt-CpG and pGFP (E). Culture supernatants were used to infect TZM-bl reporter cells to measure the amount of infectious virus production. Gag expression in the media as well as Gag, Hsp90, Env and Actin expression in the cell lysates was detected using immunoblotting (B and E). Genomic RNA abundance was quantified by qRT-PCR in cell lysates (C) and media (D). The bar charts show the average values of three (B and E) or four (C and D) independent experiments normalized to the value obtained for wild type HIV-1. Error bars represent standard deviation. *p<0.05 as determined by a two-tailed unpaired t-test comparing the virus containing synonymous mutations to wild type HIV-1.

Fig. 9. CpG dinucleotides in gag have ZAP-dependent and ZAP-independent effects but the number of CpGs does not correlate with antiviral activity. (A, C and D) HeLa Control, ZAP-ex4 and ZAP-ex6 CRISPR cells were transfected with pHIV-1, pHIV-1\textsubscript{env88-561}CpG, pHIV-1\textsubscript{gag22-378}CM or pHIV-1\textsubscript{gag22-378}CpG (A), pHIV-1, pHIV-1\textsubscript{gag22-222}CpG, pHIV-1\textsubscript{gag22-222}65CpG or pHIV-1\textsubscript{gag22-222}660CpG (C), or pHIV-1, pHIV-1\textsubscript{env88-561}CpG, pHIV-1\textsubscript{pol795-1386}CpG, pHIV-1\textsubscript{gag694-1206}CpG plus pGFP (D). Culture supernatants were used to infect TZM-bl reporter cells to measure infectious virus production. Gag expression in the media, as well as Gag, Hsp90, Env and Actin...
expression in the cell lysates was detected using immunoblotting. The bar charts show the average values of three (C and D) or four (A) independent experiments normalized to HIV-1 in HeLa Control CRISPR cells. Error bars represent standard deviation. *p<0.05 as determined by a two-tailed unpaired t-test. The black * compares the virus containing introduced CpGs in the Control CRISPR cells to wild type HIV-1 in the Control CRISPR cells. The red * compares the virus containing introduced CpGs between the ZAP CRISPR cells and the Control CRISPR cells. (B) Schematic representation of the HIV-1 gag with the region of synonymously introduced CpGs in HIV-1\textsubscript{gag22-1185}CpG, HIV-1\textsubscript{gag22-651}CpG, HIV-1\textsubscript{gag660-1185}CpG and HIV-1\textsubscript{gag694-1206}CpG highlighted.

Fig. 10. Codon modification in the 5' end of gag activates a cryptic splice donor. (A) Schematic representation of the HIV-1 gag region with the introduced synonymous mutations in HIV-1\textsubscript{gag22-165}CM, HIV-1\textsubscript{gag22-261}CM, or HIV-1\textsubscript{gag22-378}CM highlighted. (B) HeLa cells were transfected with pHIV-1, pHIV-1\textsubscript{gag22-165}CM, pHIV-1\textsubscript{gag22-261}CM, or pHIV-1\textsubscript{gag22-378}CM. Total and genomic RNA abundance were quantified by qRT-PCR in the cell lysates. The bar charts show the average values of two independent experiments normalized to wild type HIV-1. (C) Relative usage of splice donor 1 and cryptic splice donor 1 upon codon modification of HIV-1. The 9173 nt HIV-1 genomic RNA and features are depicted in the "HIV-1 genome" track. Canonical donors (D1-4) and acceptors (A1-7), codon modification-induced-cryptic donor (CD1), and gag start codon (AUG) are shown in the “Splice Sites and Gag AUG” track. The number of reads supporting use of D1 (nt 291) or CD1 (nt 718) paired with a given splice acceptor is depicted by line width (y-axis). The splice data line height is arbitrary. Read numbers shown are the sum across duplicate samples.
per thousand HIV-mapping reads. (D) Alignment of gag sequences for HIV-1, HIV-1\textsubscript{gag22-165CM}, HIV-1\textsubscript{gag22-261CM}, and HIV-1\textsubscript{gag22-378CM}. Introduced CpGs are highlighted in red and CpGs found in wild type HIV-1 gag sequence are highlighted in yellow. The cryptic splice donor (CD1) activated by the synonymous mutations in gag nt 22-378 is highlighted in green.
Table 1: CpGs and mutations introduced by codon modification

| Virus                        | Total CpGs in wild type virus within the modified region | Total CpGs introduced within the modified region | Total mutations introduced within the modified region |
|------------------------------|---------------------------------------------------------|--------------------------------------------------|------------------------------------------------------|
| HIV-1 gag22-165 CM           | 4                                                       | 11                                               | 49                                                   |
| HIV-1 gag22-261 CM           | 4                                                       | 18                                               | 80                                                   |
| HIV-1 gag22-378 CM           | 4                                                       | 26                                               | 109                                                  |
| HIV-1 gag22-378 CM-no-CpG    | 4                                                       | 0                                                | 79                                                   |
| HIV-1 gag22-378 CpG          | 4                                                       | 26                                               | 30                                                   |
| HIV-1 gag22-378 DC           | 4                                                       | 0                                                | 23                                                   |
| HIV-1 gag22-378 CM-5nt-CpG   | 4                                                       | 26                                               | 71                                                   |
| HIV-1 gag22-651 CpG          | 4                                                       | 30                                               | 37                                                   |
| HIV-1 gag660-1185 CpG        | 5                                                       | 32                                               | 36                                                   |
| HIV-1 gag22-1185 CpG         | 9                                                       | 62                                               | 73                                                   |
| HIV-1 gag694-1206 CpG        | 5                                                       | 60                                               | 72                                                   |
| HIV-1 pol795-1388 CpG        | 1                                                       | 53                                               | 61                                                   |
| HIV-1 env66-561 CpG          | 1                                                       | 36                                               | 43                                                   |
| HIV-1 env611-1014 CpG        | 4                                                       | 48                                               | 52                                                   |
| HIV-1 env66-1014 CpG         | 5                                                       | 84                                               | 95                                                   |
| HIV-1-IRES-GFP               | N/A                                                     | 96                                               | N/A                                                  |
| HIV-1-IRES-Renilla           | N/A                                                     | 96                                               | N/A                                                  |
| HIV-1 Gag-Venus              | N/A                                                     | 64                                               | N/A                                                  |
| HIV-1 Gag-STOP-Venus         | N/A                                                     | 64                                               | N/A                                                  |
Table 2: Splicing events from the splice donor 1 (D1) or cryptic splice donor 1 (CD1) to the canonical HIV-1 splice acceptors as a percentage of the total HIV-1 splicing events.

| Start | End   | Splice sites | WT.1 | WT.2 | CM22-165.1 | CM22-165.2 | CM22-261.1 | CM22-261.2 | CM22-378.1 | CM22-378.2 |
|-------|-------|--------------|------|------|------------|------------|------------|------------|------------|------------|
| 291   | 4460  | D1-A1        | 11.9 | 11.9 | 3.0        | 3.4        | 2.5        | 1.5        | 0.2        | 0.0        |
| 291   | 4937  | D1-A2        | 4.4  | 4.4  | 3.2        | 3.3        | 2.9        | 2.1        | 0.0        | 0.0        |
| 291   | 5324  | D1-A3        | 3.0  | 3.1  | 3.4        | 3.4        | 3.5        | 3.1        | 0.0        | 0.0        |
| 291   | 5483  | D1-A4c       | 1.2  | 1.4  | 0.6        | 1.1        | 0.1        | 0.4        | 0.0        | 0.0        |
| 291   | 5501  | D1-A4a       | 4.1  | 3.7  | 1.9        | 2.0        | 1.6        | 0.9        | 0.0        | 0.0        |
| 291   | 5507  | D1-A4b       | 2.2  | 2.2  | 1.7        | 1.6        | 1.1        | 1.1        | 0.3        | 0.1        |
| 291   | 5523  | D1-A5        | 25.7 | 26.2 | 30.4       | 33.3       | 29.7       | 26.2       | 0.7        | 0.9        |
| 291   | 7916  | D1-A7        | 0.9  | 1.0  | 0.9        | 0.8        | 0.0        | 0.4        | 0.0        | 0.0        |
| 718   | 4460  | CD1-A1       | 0.0  | 0.0  | 0.6        | 0.4        | 2.0        | 1.5        | 7.8        | 7.4        |
| 718   | 4937  | CD1-A2       | 0.0  | 0.0  | 0.0        | 0.0        | 0.5        | 0.8        | 1.7        | 1.4        |
| 718   | 5324  | CD1-A3       | 0.0  | 0.0  | 0.1        | 0.1        | 1.4        | 1.1        | 1.6        | 1.8        |
| 718   | 5483  | D1-A4c       | 0.0  | 0.0  | 0.0        | 0.0        | 0.0        | 0.0        | 0.0        | 0.0        |
| 718   | 5501  | CD1-A4a      | 0.0  | 0.0  | 0.0        | 0.0        | 0.0        | 0.1        | 1.0        | 1.2        |
| 718   | 5507  | CD1-A4b      | 0.0  | 0.0  | 0.0        | 0.0        | 0.4        | 0.4        | 3.2        | 3.0        |
| 718   | 5523  | CD1-A5       | 0.0  | 0.0  | 0.2        | 0.4        | 4.7        | 4.4        | 24.9       | 23.1       |
| 718   | 7916  | CD1-A7       | 0.0  | 0.0  | 0.0        | 0.0        | 0.0        | 0.0        | 0.0        | 0.0        |

1233
1234
1235
1236
1237
Figure 1

A. 

B. 

C. 

D.
Figure 6

A. 

| Treatment | Control | ZAP-ex6 | Control | ZAP-ex6 | Control | ZAP-ex6 |
|-----------|---------|---------|---------|---------|---------|---------|
| HIV-1     | -       | -       | +       | -       | -       | +       |
| env86-96CpG | +       | +       | -       | +       | -       | +       |
| poRT85-1386CpG | +       | +       | -       | +       | -       | +       |

Relative Infectious Virus Production

Virions (p24<sup>core</sup>)

Cell lysates

p55<sup>core</sup>
p24<sup>core</sup>

Hsp90

B. 

| Treatment | Control | ZAP-ex6 | Control | ZAP-ex6 |
|-----------|---------|---------|---------|---------|
| HIV-1     | -       | -       | +       | -       |
| HIV-1-HRES-GFP | +       | +       | -       | +       |
| HIV-1-HRES-Ranilla | +       | +       | -       | +       |

Relative Infectious Virus Production

Virions (p24<sup>core</sup>)

Cell lysates

p55<sup>core</sup>
p24<sup>core</sup>

Hsp90

C. 

| Treatment | ZAP-S | ZAP-L |
|-----------|-------|-------|
| IFN-I     | -     | +     |
| -         | +     | -     |
| -         | +     | +     |

Fold Expression Relative to Untreated

IFN-I

ZAP-L

ZAP-S

Actin

HEK293T

Control

ZAP-ex6

1 2 3 4

ns
Figure 8

A. 

B. 

C. 

D. 

E.
