CG dinucleotide suppression enables antiviral defence targeting non-self RNA

Matthew A. Takata1, Daniel Gonçalves-Carneiro1, Trinity M. Zang1,2, Steven J. Soll1,2, Ashley York1, Daniel Blanco-Melo1 & Paul D. Bieniasz1,2.

Vertebrate genomes exhibit marked CG suppression— that is, lower than expected numbers of 5′-CG-3′ dinucleotides. This feature is likely to be due to C-to-T mutations that have accumulated over hundreds of millions of years, driven by CG-specific DNA methyl transferases and spontaneous methyl-cytosine deamination. Many RNA viruses of vertebrates that are not substrates for DNA methyl transferases mimic the CG suppression of their hosts2-4. This property of viral genomes is unexplained4-6. Here we show, using synonymous mutagenesis, that CG suppression is essential for HIV-1 replication. The deleterious effect of CG dinucleotides on HIV-1 replication was cumulative, associated with cytoplasmic RNA depletion, and was exerted by CG dinucleotides in both translated and non-translated exonic RNA sequences. A focused screen using small inhibitory RNAs revealed that zinc-finger antiviral protein (ZAP)7 inhibited virion production by cells infected with CG-enriched HIV-1. Crucially, HIV-1 mutants containing segments whose CG content mimicked random nucleotide sequence were defective in unmanipulated cells, but replicated normally in ZAP-deficient cells. Crosslinking–immunoprecipitation–sequencing assays demonstrated that ZAP binds directly and selectively to RNA sequences containing CG dinucleotides. These findings suggest that ZAP exploits host CG suppression to identify non-self RNA. The dinucleotide composition of HIV-1, and perhaps other RNA viruses, appears to have adapted to evade this host defence.

To identify cis-acting RNA elements within the HIV-1 genome that are important for its replication, we generated a mutant HIV-1 sequence containing the maximum number of synonymous mutations in open reading frames (ORFs). Blocks of mutations (mean of around 125 mutations per block) were represented in 16 proviral plasmids (A–P) containing a gfp reporter (Fig. 1a). Mutant viruses were divided into three groups, depending on their replication properties. Group 1 mutants displayed near-normal viral replication and group 2 mutants were defective, exhibiting severe splicing defects (unpublished observations). Group 3 mutants yielded near-normal infectious titres when proviral plasmids were transfected into 293T cells and lacked an obvious splicing defect, but were defective in spreading replication assays (Fig. 1b, c, Extended Data Fig. 1a).

Mapping experiments that used derivatives of the defective group 3 mutant viruses L and M that contained mutated segments in env revealed that the replication defects of these viruses were not caused by perturbation of a single discrete element. Indeed, mutants LC, LD, LE, LF, MA, MC and MD, which contained smaller mutant segments, collectively representing all mutations in L and M, each replicated with kinetics close to those of wild-type HIV-1 (HIV-1WT) (Fig. 1a–d). Moreover, when the mutations in four replication-competent pol mutants (E–H, Fig. 1a) were combined, the resulting mutant virus (EH) was defective (Fig. 1e). Thus, HIV-1 replication defects were induced by cumulative effects of synonymous mutations in pol or env.

The HIV-1 genome is sparse in C mononucleotides8 and, like many vertebrate viruses3-4, is particularly deficient in CG dinucleotides, (Fig. 1f). Our synonymous mutagenesis coincidentally increased the CG dinucleotide content in mutant segments to a level similar to that of random sequence (Fig. 1f). We generated derivatives of mutant L, termed LCG and LOTH, respectively, containing only mutations that generated new CG dinucleotides (37 of 145 original mutations) or the 108 other mutations (Supplementary Data 1). We also generated mutants that maximized the CG or, as a further control, GC dinucleotide content in the same segment (LCG-HI and LGC-HI) (Extended Data Table 1). These proviral plasmids each yielded similar levels of infectious virus following transfection of 293T cells.

Figure 1 | Synonymous mutagenesis reveals inhibitory effects of CG dinucleotides on HIV-1 replication. a, Representation of HIV-1NHG, a human immunodeficiency virus type-1 provirus encoding gfp in place of nef, indicating synonymous mutant blocks, and corresponding phenotypes (see text). b–e, Replication of HIV-1 mutants in MT4 cells, as measured by fluorescent-activated cell sorting (FACS) enumeration of infected cells. f, Number of CG dinucleotides in a 200-nucleotide sliding window in viral and random sequences. g, Replication of HIV-1 mutants in MT4 cells, measured as in b. h, Replication of HIV-1 mutants in primary lymphocytes, measured by supernatant reverse transcriptase activity.
Figure 2 | CG dinucleotides cause depletion of cytoplasmic RNA. a, Single-cycle infectious virus yield, following infection of MT4 cells with equal titres of HIV-1WT and mutants (mean ± s.e.m., n = 3 independent experiments). b, Western blot analysis 48 h after a single-cycle infection of MT4 cells with wild-type or mutant HIV-1, representative of three experiments. c, Location of salient exons (black lines), mutated segments (red shading) and smFISH probes (green shading) in HIV-1 mRNAs. d, RT–qPCR quantitation of unspliced RNA in MT4 cells in a single-cycle infection assay (mean ± s.e.m., n = 2 or 4 independent experiments). (Extended Data Fig. 1a). However, LCG and LCG–HI were defective in MT4 cells, whereas LOTH and LGC–HI replicated with kinetics close to those of HIV-1WT (Fig. 1g). Mutants L and LCG–HI also replicated at about 100-fold lower levels than HIV-1WT and LGC–HI in primary lymphocytes (Fig. 1h, Extended Data Fig. 1b, c). Thus, suppression of CG but not GC dinucleotides appears to be essential for HIV-1 replication.

To understand the basis of replication defects in the CG-enriched HIV-1 mutants, we infected MT4 cells with equal titres of each virus in single-cycle replication experiments. Notably, cells infected with LCG or LCG–HI generated about 1,000-fold fewer infectious progeny virions than did cells infected with LOTH or LGC–HI (Fig. 2a). Infectious virion yields from EH-infected cells were similarly reduced (Extended Data Fig. 2a). Western blot analyses revealed reduced levels of viral Gag and Env proteins in cells infected with L, LCG or LGC–HI, but the same levels as HIV-1WT for cells infected with LOTH or LGC–HI (Fig. 2b). Expression of the gfp reporter that was embedded in the nef gene, and therefore expressed via an mRNA from which the L segment is removed by splicing (Fig. 2c), was equivalent for each virus (Fig. 2b, Extended Data Fig. 2b). A deficit in Gag protein levels also occurred in EH-infected cells. However, these cells generated normal levels of both Env and GFP proteins, whose spliced mRNAs lack the CG-enriched EH segment (Extended Data Fig. 2c).

Unspliced viral RNA levels in single-cycle infected MT4 cells, measured by reverse transcription–quantitative PCR (RT–qPCR), were five-to tenfold lower in cells infected with L, LCG, LCG–HI or EH but at HIV-1WT levels in cells infected with LOTH, LGC–HI, E, F, G or H (Fig. 2d, Extended Data Fig. 2d). Single-molecule fluorescence in situ hybridization (smFISH) experiments using a gag probe revealed that the deficit in unspliced viral RNA occurred specifically in the cytoplasm in LCG–HI-infected cells, whereas levels of unspliced RNA in the nucleus were equivalent for HIV-1WT, LGC–HI and LGC–HI (Fig. 2e, f, Extended Data Fig. 3). Similar smFISH experiments using a probe that detected all spliced and unspliced viral RNAs (Fig. 2e) revealed a marginal, statistically ambiguous deficit for cells infected with LCG–HI (Extended Data Figs 2e, 4). Thus, incompletely spliced RNAs (which represent only a subset of total HIV-1 RNAs) appeared to be selectively depleted in LCG–HI-infected cells.

Because a deficit in levels of CG-containing RNAs and their protein products was the foundational defect in cells infected with the defective viral mutants, we conducted a focused small inhibitory RNA (siRNA) screen targeting proteins involved in RNA degradation pathways (for example, the microRNA, nonsense-mediated decay and RNA exosome pathways) (Fig. 3a, Extended Data Fig. 5a). Single-cycle replication experiments revealed that knockdown of ZAP7 almost completely restored infectious virion yield from LCG–HI-infected cells (Fig. 3a). Knockdown of TRIM25, which enhances ZAP activity10,11, also increased viral yield.

We generated ZAP–/– MT4 cells lacking both major ZAP isoforms (ZAP-L and ZAP-S, Fig. 3b) using CRISPR–Cas9 genome editing. While previous work has indicated that an overexpressed ZAP fragment can inhibit HIV-1 infection42, knockout of endogenous ZAP in MT4 cells did not affect HIV-1WT or LGC–HI replication (Fig. 3c). Strikingly, LCG–HI and EH, which were defective in unmanipulated MT4 cells, replicated with kinetics that were indistinguishable from those of HIV-1WT in ZAP–/– MT4 cells (Fig. 3c). The deficits in viral protein levels observed in cells infected with CG-enriched viruses were abolished in ZAP–/– cells (Fig. 3d). Reconstitution of ZAP–/– MT4 cells with a CRISPR-resistant, doxycycline-inducible ZAP-S construct (ZAPD2) enabled doxycycline-dependent inhibition of CG-enriched virus replication and protein expression in single-cycle assays, but did not affect HIV-1WT (Fig. 3c, Extended Data Fig. 5b, c). Moreover, the deficit in unspliced viral RNA seen in cells infected with CG-enriched viruses was abolished in ZAP–/– cells and reinstated in a doxycycline-dependent manner in ZAPD2-reconstituted ZAP–/– cells (Fig. 3e).

We transferred the L-mutant segment and its derivatives into the 3′ UTR of a reporter construct encoding a synthetic CG-depleted fluc gene (Extended Data Fig. 6a). The CG-enriched L-derived elements inhibited luciferase expression approximately fivefold in this context in a simple plasmid transfection assay (Fig. 3f). These inhibitory effects were abolished when ZAP–/– HeLa cells were transfected (Fig. 3f, Extended Data Fig. 6b). Elevating the CG dinucleotide content of fragments from naturally CG-suppressed vesicular stomatitis virus or influenza virus-derived RNA sequences43 (Extended Data Fig. 6a, c) conferred sensitivity to ZAP-L in cotransfection assays with similar reporter constructs (Extended Data Fig. 6d).
ZAP has been reported to bind RNA, but no shared features of its target sequences are evident. To determine the RNA binding specificity of ZAP, we used crosslinking–immunoprecipitation–sequencing (CLIP–seq) assays in cells infected with HIV-1WT or mutant L. Remarkably, ZAP bound to the HIV-1 genome predominantly at the precise location of the CG-enriched segment in mutant L (Fig. 4a, b). Conversely, ZAP bound less frequently to HIV-1WT and the unaltered portions of the L genome. Inefrequent ZAP binding sites in the HIV-1 genome almost always coincided with rare CG dinucleotides (Fig. 4a, Extended Data Fig. 7a).

Although the L mutant genome was the single most frequently bound RNA in infected cells, ZAP also bound cellular mRNAs (Extended Data Fig. 7b). CG suppression is marked in human mRNA ORF and 3′ UTR sequences but was absent in the subset of these sequences that represented preferred ZAP binding sites (Extended Data Fig. 7c–e). A more detailed analysis of dinucleotides that are underrepresented (CG and UA) or overrepresented (UG) in ORFs and 3′ UTRs as well as an inverted control dinucleotide (GC) revealed that ZAP binding sites were highly CG-enriched (Fig. 4c, Extended Data Fig. 7f). Conversely, UA, UG or GC dinucleotides were present in preferred ZAP binding elements at frequencies similar to those of ORFs and 3′ UTRs (Fig. 4c). A control RNA binding protein (APOBEC3G) showed no preference for CG-enriched elements (Extended Data Fig. 7g).

The replication of several, but not all, RNA and reverse-transcribing viruses is inhibited by overexpressed or endogenous ZAP. Inspection of the dinucleotide composition of these viral genomes revealed that the degree of CG suppression generally predicted ZAP resistance (Extended Data Fig. 8a). Moreover, the degree to which previously mapped viral RNA elements conferred sensitivity to ZAP in reporter assays was also generally predicted by the product of their length and the degree to which CG nucleotides were suppressed (Extended Data Fig. 8b).

CG suppression in ORFs is synonymous with codon-pair bias. However, the CG- and ZAP-dependent inhibition of HIV-1 protein expression occurred when CG-enriched elements were incorporated into exonic (but not intronic) translated or 3′ UTR portions of the corresponding pre-mRNAs. Thus, CG dinucleotides exert effects post-transcriptionally but independent of codon-pair translation efficiency. Rather, direct ZAP recognition causes cytoplasmic depletion of RNAs with high CG content. While ZAP can also regulate the levels...
of certain host mRNAs (for example, that encoding TRAILR4)\(^{28}\), this activity requires the C-terminal poly ADP ribose polymerase (PARP) domain that is absent in ZAP-S (which exhibits antiviral activity). Most cellular mRNAs are unaffected by ZAP\(^{28}\). Thus, it appears that the main targets of ZAP are non-self, viral RNAs in which CG suppression is incomplete. Manipulation of the CG content in viruses\(^{25,26}\) and manipulation of ZAP in cells could enable levels of viral attenuation or recombinant gene expression to be adjusted, with many possible applications.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 24 May; accepted 30 August 2017.**

**Published online 27 September 2017.**

1. Karlin, S. & Mrázek, J. Compositional differences within and between eukaryotic genomes. *Proc. Natl Acad. Sci. USA* **94**, 10227–10232 (1997).
2. Karlin, S., Doerfler, W. & Cardon, L. R. 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 (1994).
3. Rima, B. K. & McFerran, N. V. Dinucleotide and stop codon frequencies in single-stranded RNA viruses. *J. Gen. Virol.* **78**, 2859–2870 (1997).
4. Greenbaum, B. D., Levine, A. J., Bhanot, G. & Rabadán, R. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog.* **4**, e1000079 (2008).
5. Chang, X. et al. CpG usage in RNA viruses: data and hypotheses. *PLoS One* **8**, e74109 (2013).
6. Futcher, B. et al. Reply to Simmonds et al.: Codon pair and dinucleotide bias have not been functionally distinguished. *Proc. Natl Acad. Sci. USA* **112**, E3635–E3636 (2015).
7. Gao, G., Guo, X. & Goff, S. P. Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein. *Science* **297**, 1703–1706 (2002).
8. van Hemert, F., van der Kuyl, A. C. & Berkhout, B. On the nucleotide composition and structure of retroviral RNA genomes. *Viruses* **13**, 16–23 (2014).
9. Karin, J. & Stoltzfus, C. M. Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harb. Perspect. Med.* **2**, a006916 (2012).
10. Li, M. M. et al. TRIM25 enhances the antiviral action of zinc-finger antiviral protein (ZAP). *PLoS Pathog.* **13**, e1006145 (2017).
11. Zheng, X. et al. TRIM25 is required for the antiviral activity of zinc finger antiviral protein. *J. Virol.* **91**, e00388–17 (2017).
12. Zhu, Y. et al. Zinc-finger antiviral protein inhibits HIV-1 infection by selectively targeting multiply spliced viral mRNAs for degradation. *Proc. Natl Acad. Sci. USA* **108**, 15834–15839 (2011).
13. Guo, X., Carroll, J. W., Macdonald, M. R., Goff, S. P. & Gao, G. The zinc finger antiviral protein directly binds to specific viral mRNAs through the CCCH zinc finger motifs. *J. Virol.* **78**, 12781–12787 (2004).
14. Zhu, Y. & Gao, G. ZAP-mediated mRNA degradation. *RNA Biol.* **5**, 65–67 (2008).
15. Chen, S. et al. Structure of N-terminal domain of ZAP indicates how a zinc-finger protein recognizes complex RNA. *Nat. Struct. Mol. Biol.* **19**, 430–435 (2012).
16. Huang, Z., Wang, X. & Gao, G. Analyses of SELEX-derived ZAP-binding RNA aptamers suggest that the binding specificity is determined by both structure and sequence of the RNA. *Protein Cell* **1**, 752–759 (2010).
17. Bick, M. J. et al. Expression of the zinc-finger antiviral protein inhibits alphavirus replication. *J. Virol.* **77**, 65–70 (2003).
18. Müller, S. et al. Inhibition of filovirus replication by the zinc finger antiviral protein. *J. Virol.* **81**, 2391–2400 (2007).
19. Mao, R. et al. Inhibition of hepatitis B virus replication by the host zinc finger antiviral protein. *PLoS Pathog.* **9**, e1003494 (2013).
20. Lin, Y. et al. Identification and characterization of alphanavirus M1 as a selective oncolytic virus targeting ZAP-defective human cancers. *Proc. Natl Acad. Sci. USA* **111**, E4504–E4512 (2014).
21. Goodier, J. L., Pereira, G. C., Cheung, L. E., Rose, R. J. & Kazazian, H. H., Jr. The broad-spectrum antiviral protein ZAP restricts human retrotransposition. *PLoS Genet.* **11**, e1005121 (2015).
22. Liu, C. H., Zhou, L., Chen, G. & Krug, R. M. Battle between influenza A virus and a newly identified antiviral activity of the PARP-containing ZAP protein. *Proc. Natl Acad. Sci. USA* **112**, 14048–14053 (2015).
23. Tang, Q., Wang, X. & Gao, G. The short form of the zinc finger antiviral protein inhibits influenza A virus protein expression and is antagonized by the virus-encoded NS1. *J. Virol.* **91**, e01909–16 (2017).
24. Coleman, J. R. et al. Virus attenuation by genome-scale changes in codon pair bias. *Science* **320**, 1784–1787 (2008).
25. Tulloch, F., Atkinson, N. J., Evans, D. J., Ryan, M. D. & Simmonds, P. RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. *eLife* **3**, e04531 (2014).
26. Kunec, D. & Osterrieder, N. Codon pair bias is a direct consequence of dinucleotide bias. *Cell Host Microbe* **17**, 155–167 (2015).
27. Todorova, T., Bock, F. J. & Chang, P. PARP13 regulates cellular mRNA post-transcriptionally and functions as a pro-apoptotic factor by destabilizing TRAILR4 transcript. *Nat. Commun.* **5**, 5362 (2014).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank T. Kueck for primary lymphocytes and S. Giese for assistance with smFISH. This work was supported NIH grants R01AI50111 and P50GM103297 (to D.G-C.).

**Author Contributions** M.A.T. performed all experiments unless otherwise stated and wrote the paper. D.G.-C. performed some of the luciferase reporter experiments and bioinformatic analyses. T.M.Z. performed smFISH for assistance with smFISH. This work was supported NIH grants R01AI50111 and P50GM103297 (to D.G.-C.).

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to P.D.B. (pbeniasz@rockefeller.edu).

**Reviewer Information** Nature thanks G. Towers and the other anonymous reviewer(s) for their contribution to the peer review of this work.
METHODS

Plasmid constructs. A synonymously mutated HIV-1 sequence was designed that contained a maximum number of substitutions in open reading frames. Mutations were designed to maximize the probability of disrupting secondary structure by incorporating transversion mutations (purine to pyrimidine or vice versa) where possible. No new AG or GU dinucleotides were introduced, to avoid the creation of new splice acceptors and donors. Sequences encoding overlapping open reading frames were not altered, and all known cis-acting elements that control HIV-1 splicing, gene dosage, and other aspects of HIV-1 replication were intact in the mutant viral genome.

This designed HIV-1 sequence contained 1,976 synonymous mutations. It was divided into 150–500-nucleotide blocks (A–P), which were synthesized (Genewiz) and introduced in place of native sequence into HIV-1NL4-3, a proviral plasmid containing a reporter GFP embedded in nef, or HIV-1NL4-3.L, using restriction sites proximal to the mutated regions. Supplementary Data 1 contains a codon-by-codon list of the mutations made in segment L, and Supplementary Data 2 and 3 contain alignments of the wild-type and mutant HIV-1 NL4-3 L segments (Fasta format). A complete characterization of the virus mutants not described in detail in this manuscript will be published elsewhere (M.T., S.J.S., D.R.M. and P.D.R.).

Cells were activated with phytohaemagglutinin (Sigma, 5 μg ml−1) in RPMI with 10% fetal bovine serum. Cells were activated with phytohaemagglutinin (Sigma, 5 μg ml−1) in RPMI with 10% fetal bovine serum.

Western blot analyses and antibodies. Cells were counted, normalized for cell number, lysed in SDS sample buffer, separated by electrophoresis on NuPage 4–12% Bis-Tris gels (Novex) and blotted onto nitrocellulose membranes (GE Healthcare). Antibodies against PTB1 (ab5642), Drosha (ab12286), DICR (ab4601), EXOC6 (ab50916), EXOC5 (ab11455) and PARN (ab188333) were obtained from Abcam. Antibodies for Upf1 (A300-036A), METTL3 (A301-567A), EXOSC4 (A303-775A), EXOSC5 (A303-887A), and Xrn1 (A300-443A) were obtained from Proteintech. The HIV-1 capsid antibody (183-H12-5C) was obtained from the NIH (A303-775A), EXOSC5 (A303-887A), and Xrn1 (A300-443A) were obtained from Proteintech. The HIV-1 capsid antibody (183-H12-5C) was obtained from the NIH.

For single cycle replication, MT4 cells were infected at an MOI of 1.0 with HIV-1NL4-3-derived viruses, washed three times with PBS 18 h after infection, and resuspended in RPMI with 50 μg ml−1 dextran sulphate to prevent reinfection. At 48 h after infection, cells and supernatants were collected for analysis. Half of the cells were lysed in SDS sample buffer for Western blot analysis and half allocated for RNA extraction and to determine levels of unspliced RNA as described below. The supernatants were filtered with a 0.22-μm filter. An aliquot of filtered supernatant was used to determine infectious virion yield by titration on MT4 cells. The remaining supernatant was centrifuged at 20,000 g for 1 h at 4 °C and stored at −80 °C.

RNA interference screen. Cells were transfected with 50 pmol siRNA SMART pool (Dharmacon) using RNAiMAX (Thermo Fisher) in a 6-well plate seeded in primary lymphocyte replication and CLIP assays were generated by transfection with HIV-1NL4-3. Viruses used to infect CD4-negative HeLa cells in the single cycle replication siRNA screen, or 293T cells in the CLIP assays, were generated by transfection with 10 μg proviral plasmid and 1 μg VSV-G expression plasmid.

Infection assays. Titres of viral stocks were determined by performing threefold serial dilutions in a 96-well plate and infecting 5 × 104 MT4 cells per well. At 16–18 h after infection, dextran sulphate was added to each well at a concentration of 50 μg ml−1 to prevent reinfection by nascent virions. At 48 h after infection, cells were washed twice in 4% PFA and enumerated by FACS analysis using a CyFlow Space cytometer (Partec) coupled to a Hypercyte Autosampler (Intelligent).

For spreading replication assays with GFP reporter viruses, viral stocks generated from transfected 293T cells were adjusted to the same number of single cycle infectious units (determined on MT4 cells as described above). Thereafter, 2 × 104 MT4 cells were infected at an MOI of 0.002 in 2 ml RPMI. Aliquots of infected cells were withdrawn each day and fixed in 4% PFA, and the proportion of infected cells was determined by FACS analysis of GFP expression. For spreading replication assays in PBMCs, cells were infected at an MOI of 0.001. At 18 h after infection, the cells were washed four times with PBS and cultured in RPMI with 50 U ml−1 interleukin-2. Supernatants were collected every 24 h. Viral particle release was determined by measuring the reverse transcriptase activity with a PCR-based assay.

For single cycle replication, MT4 cells were infected at an MOI of 1.0 with HIV-1NL4-3-derived viruses, washed three times with PBS 18 h after infection, and resuspended in RPMI with 50 μg ml−1 dextran sulphate to prevent reinfection. At 48 h after infection, cells and supernatants were collected for analysis. Half of the cells were lysed in SDS sample buffer for Western blot analysis and half allocated for RNA extraction and to determine levels of unspliced RNA as described below. The supernatants were filtered with a 0.22-μm filter. An aliquot of filtered supernatant was used to determine infectious virion yield by titration on MT4 cells. The remaining supernatant was centrifuged at 20,000 g for 1 h at 4 °C and stored at −80 °C.
with $2 \times 10^5$ HeLa cells per well. At 24 h after transfection, the cells were infected with either wild-type HIV-1\textsubscript{NHG} or L\textsubscript{CG-HE}. At 48 h after transfection, the cells were washed three times with PBS and suspended in DMEM. The cells and supernatant were collected 72 h after transfection to determine knockdown efficiency levels and the yield of infectious virions.

**CLIP-seq.** The CLIP method used in this study has been previously described\(^{29}\). In brief, RNA and proteins were cross linked by feeding cells overnight with 4-thiouridine and irradiating them at 0.15 J cm\(^{-2}\) UV ($\lambda = 365$ nm) in a Stratalinker 2400 UV crosslinker (Stratagene). Thereafter, ZAP-3\times HA was immunopurified using Protein G-conjugated magnetic Dynabeads and a mouse monoclonal anti-HA antibody, and the RNA was radiolabelled with $0.5 \mu$Ci ml\(^{-1}\) $\gamma$-32P[ATP]. Protein–RNA adducts were separated by SDS–PAGE, transferred to nitrocellulose and detected by autoradiography. Next, sequential 3\' and 5\' adaptor ligations were performed as previously described\(^{29}\) by attaching a known sequence that contains primer binding sites for reverse transcription and PCR-amplification of the cDNA library. Sequencing of the cDNA library was done on an Illumina HiSeq 2000 platform.

The analysis pipeline used in this study has previously been described\(^{29}\). Processing of raw reads was performed with the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), excluding reads with fewer than 15 nucleotides. Reads were then aligned to the human genome (hg38) concatenated with the HIV-1\textsubscript{NL4-3} genome or to the viral genome alone. Cluster analysis was performed using PARalyzer\(^{30}\).

**Statistical information.** HIV-1 replication experiments were done at least three times and representative data are shown. CLIP experiments were done four times and representative data are shown. Statistical analysis of smFISH data was done using the Mann–Whitney $U$ test. Statistical analysis of variation in dinucleotide frequencies between ORFs, 3\' UTRs and CLIP-derived ZAP-binding sequences and the human databases of ORFs and 3\' UTRs were performed using Welch's unequal variance $t$-test implemented using R. All other experiments yielding quantitative data were done at least three times and mean values ± s.e.m. are plotted. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

**Data availability.** The sequence of the NHG proviral plasmid is available at Genbank (accession code JQ585717.1), as is the NHG-derived sequence containing all 12 mutant segments (accession code MF944225). The HIV-1 L mutant sequence has also been submitted to GenBank (accession code MF687717 the CLIP–seq data have been deposited in the NCBI GEO data repository with accession code GSE102843 (GSM2747099 and GSM2747100).

29. Kutluay, S. B. et al. Global changes in the RNA binding specificity of HIV-1 gag regulate virion genesis. *Cell* **159**, 1096–1109 (2014).

30. Corcoran, D. L. et al. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.* **12**, R79 (2011).
Extended Data Figure 1 | CG-enriched HIV-1 clones yield near wild-type levels of virus from transfected 293T cells but are attenuated in replication in primary lymphocytes. a, Yield of infectious virus from proviral plasmid transfected 293T cells, as measured by infection of MT4 cells (mean ± s.e.m., n = 3, 4 or 5 independent experiments). b, c, Spreading replication of HIV-1 mutants in primary lymphocytes from two additional donors as measured by reverse transcriptase activity in the supernatant of infected cells over time.
Extended Data Figure 2 | Effects of CG dinucleotides on HIV-1 infectious virion yield, RNA and protein levels in single-cycle replication assays. 

**a**, Yield of infectious virus in a single cycle of replication following infection of MT4 cells with equal titres of HIV-1WT and pol mutants (mean ± s.e.m., n = 3 independent experiments).

**b**, Expression of gfp in MT4 cells, as measured by flow cytometry, 48 h after infection with equal titres of the indicated viruses. Numerical values are mean fluorescent intensity (MFI) of infected cells (indicated by the dotted box).

**c**, Western blot analysis (anti-Gag, anti-Env, anti-GFP and anti-HSP90) of viral, reporter and cellular protein expression, 48 h after a single cycle of infection of MT4 cells with wild-type and synonymous pol mutant HIV-1. Representative of three experiments.

**d**, RT–qPCR quantification of unspliced RNA in MT4 cells in a single-cycle infection assay with wild-type and synonymous pol mutant HIV-1 (mean ± s.e.m., n = 2 or 3 independent experiments).

**e**, Quantification of RNA molecules (fluorescent spots) by smFISH in cytoplasm using a probe targeting all spliced and unspliced HIV-1 RNA species after infection of HOS/CXCR4-CD4 cells. Each symbol represents an individual cell. Horizontal lines represent mean values; P values were determined using Mann–Whitney test (n = 10).
Extended Data Figure 3 | smFISH quantification of unspliced HIV-1 RNA in infected cells. Examples of smFISH analysis of wild-type and synonymous mutant HIV-1-infected cells (red, smFISH gag probe (see Fig. 2c); green, GFP; blue, Hoescht dye). The boxed areas indicate regions selected for expanded views in Fig. 2f. Clusters of RNA molecules in the nuclei of some infected cells may represent sites of proviral integration. Representative of three independent experiments. Scale bar, 5 μm.
Extended Data Figure 4 | smFISH quantification of total HIV-1 RNA in infected cells. Examples of smFISH analysis of wild-type and synonymous mutant HIV-1-infected cells (red, smFISH probe targeting all viral mRNA species (see Fig. 2c); green, GFP; blue, Hoescht dye). Clusters of RNA molecules in the nuclei of some infected cells may represent sites of proviral integration. Representative of three independent experiments. Scale bar, 5 μm.
Extended Data Figure 5 | ZAP mediates deleterious effects of CG dinucleotides on HIV-1 replication. a, Western blot analyses, using the indicated antibodies, following transfection of HeLa cells with the corresponding siRNAs, or control siRNAs, in the single-cycle replication assays described in Fig. 3a. Representative of 2 experiments. b, Western blot analysis of ZAP expression in control, CRISPR-knockout MT4 cells and doxycycline-inducible ZAP-S-reconstituted MT4 cells. Asterisks indicate protein species that appeared in some CRISPR knockout clones, reacted with an anti-ZAP antibody and arose after extended passage. These are likely to represent truncated forms of ZAP-L whose translation initiated at methionine codons 3’ to the CRISPR target site (near the ZAP N terminus). Representative of three experiments. c, Western blot analysis (anti-Gag, anti-Env, anti-GFP and anti-tubulin) of viral and cellular protein levels in cells and virions, 48 h after single-cycle wild-type or mutant HIV-1 infection of ZAP \( ^{-/} \) MT4 cells that had been reconstituted with a doxycycline-inducible ZAP-S expression construct (ZAP\(_{S}^{(+)}\)) and left untreated or treated with doxycycline. Representative of three experiments.
**Extended Data Figure 6 | CG dinucleotides in 3′ UTRs confer sensitivity to inhibition by ZAP.**

**a.** Schematic representation of a reporter construct encoding a CG dinucleotide-depleted fluc cDNA into which were inserted the indicated sequences as 3′ UTRs. **b.** Western blot analysis of ZAP expression following CRISPR mutation of ZAP exon 1 in HeLa cells. Representative of three experiments. **c.** Number of CG dinucleotides present in a 200-nucleotide sliding window in the indicated viral cDNA sequences that were left unmanipulated (WT) or recoded with synonymous mutations to contain the maximum number of CG dinucleotides (CG+). **d.** Luciferase expression following transfection of 293T ZAP−/− cells with CG dinucleotide-depleted fluc reporter plasmids incorporating the indicated VSV or influenza A virus (IAV) RNA sequences as 3′ UTRs, in the presence or absence of a cotransfected ZAP-L expression plasmid (mean ± s.e.m., n = 4 independent experiments).
Extended Data Figure 7 | Dinucleotide composition of ORFs, 3′ UTRs, and preferred ZAP binding sites in cellular mRNAs. **a,** Expanded views of the portion of the CLIP graphs in Fig. 4a corresponding to unmutated portions of the viral genome. **b,** Sources of RNA reads bound to ZAP in a typical CLIP–seq experiment, done using HIV-1-infected cells. **c–e,** Ratio of the observed frequency to the expected frequency (obs/exp, based on mononucleotide composition) for each of the 16 possible dinucleotides, in ORFs (c), 3′ UTR sequences (d) and the 100 sites in cellular mRNAs that were most frequently bound by ZAP, based on CLIP read numbers (e). **f,** Frequency distributions of CG dinucleotide observed/expected frequencies in human ORFs, 3′ UTRs and top 100, top 1,000 and top 10,000 ZAP-binding sites in CLIP experiments. The top 100, top 1,000 and top 10,000 ZAP-binding sites account for 6.7%, 18.9% and 46.7% of total reads. **g,** Frequency distributions of CG, GC, UA and UG dinucleotide observed/expected frequencies in human ORFs, 3′ UTRs and the top 100 APOBEC3G-binding sites in CLIP assays.

Plotted values are mean ± s.d. of all ORFs (n = 35,170) and 3′ UTRs (n = 135,557) in the respective libraries or the most preferred ZAP binding sites (n = 100).
Extended Data Figure 8 | Analysis of CG suppression in previously reported ZAP-sensitive and ZAP-resistant viruses and ZAP-sensitizing elements. a, CG suppression in RNA and reverse transcribing viruses previously reported to be ZAP sensitive (n = 9, open symbols) and ZAP resistant (n = 4, filled symbols)\(^{7,17-20}\). The viruses included in the analysis and their degrees of CG suppression (CG observed/expected) are: ZAP-sensitive: Sinbis virus (0.90), Semliki forest virus (0.89), Venezuelan equine encephalitis virus (0.76), ebolavirus (0.60), hepatitis B virus (0.52), Moloney murine leukaemia virus (0.51), Marburg virus (0.53), alphavirus M1 (0.89), Ross river virus (0.82); ZAP-insensitive: HIV-1 (0.21), yellow fever virus (0.38), vesicular stomatitis virus (0.48), poliovirus (0.54). The \(P\) value was calculated using Student’s \(t\)-test (two-sided, \(n = 9\) ZAP-sensitive viruses and \(n = 4\) ZAP-resistant viruses).

Influenza virus (CG obs/exp = 0.44), which has been reported to be ZAP-resistant owing to the presence of an antagonist\(^{24}\) and ZAP-L-sensitive via an entirely distinct protein interaction-based mechanism\(^{23}\), was excluded from this analysis. b, Analysis of previous published data on ZAP inhibition of reporter gene expression. Each RNA element derived from the indicated RNA viruses was placed in a 3′ UTR of a luciferase reporter plasmid and fold inhibition by coexpressed ZAP is plotted against the product of CG suppression (CG observed/expected) and length for each RNA element. A data point that is a quantitative outlier from the general trend (indicated in red) is from the Sinbis (SINV) genome, but is nevertheless included in the linear regression analysis. \(P\) value was calculated using the \(F\)-test (two-sided, \(n = 32\) data points) Data are from refs 13 and 18.
Extended Data Table 1 | Mutations in the HIV-1 L mutant and its derivatives

| Virus     | Mutations                                      | CG dinucleotides added | CG dinucleotides total | GC dinucleotides total |
|-----------|------------------------------------------------|------------------------|------------------------|------------------------|
| WT        | none                                           | 0                      | 2                      | 18                     |
| L         | 145 synonymous mutations                       | 37                     | 39                     | 25                     |
| L<sub>CG</sub> | 37 mutations (subset of L mutations that generate new CG dinucleotides) | 37                     | 39                     | 27                     |
| L<sub>OTH</sub> | 108 mutations (subset of L mutations that do not generate new CG dinucleotides) | 0                      | 2                      | 19                     |
| L<sub>CG-HI</sub> | 41 mutations                                   | 41                     | 43                     | 18                     |
| L<sub>GC-HI</sub> | 9 mutations                                    | 0                      | 2                      | 27                     |
# Experimental design

1. **Sample size**
   - Describe how sample size was determined.
   
   Sample sizes were determined based on the Authors’ experience of what is necessary to generate a convincing and compelling result. Effect sizes in the experiments are such that any reasonable observer would conclude that the sample sizes employed in the manuscript are sufficient to support the conclusions.

2. **Data exclusions**
   - Describe any data exclusions.
   
   Some Q-RT-PCR data were excluded due to a PCR machine breakdown that was detected only after the experiment was completed. (trends were the same as those reported, but raw values were ~ 100-fold lower than those reported.) Some CLIP experiments with low read counts were excluded (see below).

3. **Replication**
   - Describe whether the experimental findings were reliably reproduced.
   
   The findings in this paper were remarkably reproducible. Every experiment was performed multiple times with essentially the same result. One exception is CLIP (Figure 4) - This is an inherently complex, long experiment. Occasionally sample loss, or reagent deterioration can result in low read counts. On such occasions, the entire experiment is discarded and repeated. During the course of this study we noticed read counts were declining, so we replaced all reagents, which fixed the problem. The CLIP results shown in figure 4 have been obtained on multiple occasions and by two different operators. We are absolutely confident in their veracity.

4. **Randomization**
   - Describe how samples/organisms/participants were allocated into experimental groups.
   
   N/A

5. **Blinding**
   - Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   
   No blinding was done in this study. Virtually all the data are quantitative, most measurements are made using a machine, and not easily subject to operator bias.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.
6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

| Item                                                                 | Status |
|----------------------------------------------------------------------|--------|
| The exact sample size ($n$) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) | Confirmed |
| A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly. | Confirmed |
| A statement indicating how many times each experiment was replicated | Confirmed |
| The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) | Confirmed |
| A description of any assumptions or corrections, such as an adjustment for multiple comparisons | Confirmed |
| The test results (e.g. $p$ values) given as exact values whenever possible and with confidence intervals noted | Confirmed |
| A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range) | Confirmed |
| Clearly defined error bars | Confirmed |

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study. PRISM and R were used for statistical analysis, SSE (http://www.virus-evolution.org/Downloads/Software/) was used for analysis and manipulation of dinucleotide content. Processing of raw reads in CLIP experiments was performed with the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Cluster analysis was performed using PARalyzer. A simple custom Python script was used to count CG dinucleotides in a sliding window (Fig 1f), it is available on request.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The Nature Methods guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No restrictions other than UBMTA
9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

The following antibodies were validated through siRNA knockdown to confirm the specificity:

| Antibody | Supplier | Cat #   |
|----------|----------|---------|
| PTBP1    | Abcam    | ab5642  |
| Drosha   | Abcam    | ab12286 |
| DICR     | Abcam    | ab14601 |
| EXOSC6   | Abcam    | ab50910 |
| EXOSC10  | Abcam    | ab50558 |
| PARN     | Abcam    | ab188333|
| Upf1     | Bethyl Labs | A300-036A |
| METTL3   | Bethyl Labs | A301-567A |
| EXOSC4   | Bethyl Labs | A303-775A |
| EXOSC5   | Bethyl Labs | A303-887A |
| Xrn1     | Bethyl Labs | A300-443A |
| ZAP      | Proteintech | 16820-1-AP |

The following antibodies were validated in our laboratory using western blot assays on human cells (infected with HIV-1 where appropriate) and are cited in our labs previous publications:

| Antibody | Supplier                  | Cat #   |
|----------|---------------------------|---------|
| HIV-1 CA | NIH AIDS Repository       | 183-H12-SC |
| GFP      | Sigma                     | G1546   |
| HIV-1 gp120 | American Research Products | 12-6205-1 |
| HA       | Biolegend                 | HA.11   |
| HSP90 Rabbit | Santa Cruz             | SC-7947 |
| HSP90 Mouse | Santa Cruz             | SC-69703 |

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

b. Describe the method of cell line authentication used.

c. Report whether the cell lines were tested for mycoplasma contamination.

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

N/A