Silent codon positions in the A-rich HIV RNA genome that do not easily become A: Restrictions imposed by the RNA sequence and structure

Ben Berkhou1,2,× and Formijn J. van Hemert1

1Laboratory of Experimental Virology, Department of Medical Microbiology, Amsterdam University Medical Centers, University of Amsterdam, Meibergdreef 15, 1105AZ Amsterdam, The Netherlands and 2Department of Life and Environmental Sciences, University of Cagliari, Via Università 40, 09124 Cagliari, Montserrat, Italy

×https://orcid.org/0000-0001-9137-542X

*Corresponding author. E-mail: b.berkhout@amsterdamumc.nl

Abstract

There is a strong evolutionary tendency of the human immunodeficiency virus (HIV) to accumulate A nucleotides in its RNA genome, resulting in a mere 40 per cent A count. This A bias is especially dominant for the so-called silent codon positions where any nucleotide can be present without changing the encoded protein. However, particular silent codon positions in HIV RNA refrain from becoming A, which became apparent upon genome analysis of many virus isolates. We analyzed these 'noA' genome positions to reveal the underlying reason for their inability to facilitate the A nucleotide. We propose that local RNA structure requirements can explain the absence of A at these sites. Thus, noA sites may be prominently involved in the correct folding of the viral RNA. Turning things around, the presence of multiple clustered noA sites may reveal the presence of important sequence and/or structural elements in the HIV RNA genome.

Key words: HIV-1 evolution; RNA genome; nucleotide composition; silent codon changes; A rejection; A enrichment; mutagenesis; phylogeny.

Introduction

We documented the presence of a profound bias for the A nucleotide in the RNA genome of the human immunodeficiency virus (HIV) (Berkhout and van Hemert 1994; van Hemert and Berkhou 1995; van Hemert, van der Kuyl, and Berkhou 2013). HIV is a retrovirus that reverse transcribes its RNA genome into a double-stranded DNA form that integrates into the host cell DNA and subsequently transcribes into RNA to produce new infectious virus particles. This ‘A pressure’ accounts for approximately 40 per cent of A nucleotides present in the viral RNA genome, which occurs at the expense of the other three nucleotides. The HIV gag, pol, and env genes that encode the major viral proteins and the nef gene accumulated A nucleotides to values as high as 52.5 per cent, especially at the synonymous or ‘silent’ third codon positions, which include the 4-fold degenerate codons encoding the amino acids Val, Pro, Thr, Ala, and Gly. The A count decreases at the tat–rev gene overlap, which obviously limits the number of truly silent codon positions (Berkhout and van Hemert 1994; van Hemert and Berkhou 1995; van Hemert, van der Kuyl, and Berkhou 2013). This biased HIV nucleotide composition represents a surprisingly constant factor in an otherwise highly variable viral genome (van der Kuyl and Berkhou 2012). Similar examples of a biased nucleotide composition are present in the genomes of other viruses. Recent evidence indicates that this virus-specific property is shaped in part by the action of antiviral mechanisms of the host cell, including the interferon system and viral restriction factors (Ficarelli et al. 2020; Takata et al. 2017; Fros et al. 2017; Vabret, Bhardwaj, and Greenbaum 2017; Vabret et al. 2014).

Despite this notable A pressure in HIV, we recently recognized that particular silent codon positions in the HIV genome remain devoid of the A nucleotide, even when the genomes of very many virus isolates were analyzed. This study focuses on this minority of silent HIV genome positions that do not select the A nucleotide in at least 99 per cent of the virus isolates. We will term these particular genome positions the ‘noA’ sites and discuss the possible underlying reasons for their persistence in light of the profound evolutionary A pressure. We systematically screened the HIV RNA genome for noA sites by comparison of a large number of virus isolates and analyzed the characteristics of these noA sites in more detail. We identified 190 noA sites in a 9 kb HIV RNA genome, which represents a minority of the total number of silent codon positions. A priori, two possibilities for the presence of noA sites can be proposed. First, the presence of an important RNA sequence motif, e.g. involved in the regulation of viral RNA splicing, may prohibit the selection of A nucleotides at distinct positions in the viral genome. In other words, analysis...
of noA sites—especially when multiple noA sites are clustered—may help us to identify novel sequence motifs that play a role in virus replication. Second, an alternative possibility is that local RNA structure requirements cause the presence of noA positions or—in other words—the absence of A nucleotides. Our current in silico analyses indicate that the local RNA structure requirement is indeed a likely cause of the evolutionary conservation of noA sites in the HIV RNA genome.

Materials and methods

The viral gag, pol, env, and nef open reading frames (ORFs) were downloaded from the Los Alamos HIV database (Foley et al. 2013). HIV sequences with degenerated base codes were removed. Alignments of 1026 gag, 607 pol, 1016 env, and 1291 nef sequences were obtained (Kumar, Stecher, and Tamura 2016), and a consensus sequence was constructed for these four ORFs (Hall 1999). The numbering of nucleotides, the SHAPE-determined HIV RNA structure of the NL4-3 virus isolate, and the derived values of the nucleotide pairing probability were taken from Watts et al. (2009). Secondary RNA structure prediction was performed by means of the Mfold algorithm (Zuker 2003).

We defined noA sites as follows. They have to fulfill four criteria: (1) a third codon position in an ORF without a gene overlap; (2) a synonymous codon, thus in principle allowing synonymous substitution into A; (3) the striking absence of A is observed in at least 99 per cent of virus isolates upon alignment of the RNA genomes; and (4) alignment gaps at a candidate noA site should be restricted to less than 5 per cent of the sequences analyzed. We mutated noA sites into A in silico to create ‘A mutants’ and to score the effect on the thermodynamic stability of local RNA structures. All calculations were performed in Excel.

Results

We defined noA sites as positions in protein-coding parts of the HIV RNA genome that could change to A without affecting the encoded protein (2/3/4/6-fold degenerate codons), yet avoid this sequence polymorphism despite a wealth of genetic variation among virus isolates and a generally strong preponderance of the A nucleotide in HIV genomes. Selection criteria for noA sites are mentioned in the Materials and Methods, but in short, a third codon position allowing A, 99 per cent conservation of A absence, and less than 5 per cent gaps in the alignments. The very few sequences with an A nucleotide present at such bona fide noA sites may represent exceptions to the noA rule, but may also be due to sequencing errors and/or inaccurate sequence alignment. We restricted the search for noA sites to four viral genes (gag, pol, env, and nef) in HIV-1 subtype B genomes. The other genes (vif, vpr, vpu, tat, and rev) are relatively small and/or play a role in virus replication. Second, an alternative possibility is that local RNA structure requirements cause the presence of noA positions that should theoretically allow genetic variation without affecting the encoded protein. Of the total of 2,290 silent third base codon positions in these ORFs, a minority of only 190 sites (8.3 per cent) could be labeled as ‘noA’ sites (Table 1). Additionally, ten noA sites were present at first codon positions (3 in gag, 2 in pol, 2 in env, and 3 in nef), representing the synonymous Arg codons AGA and CGA, which were excluded from this analysis that focuses exclusively on third codon positions. Silent noA sites at the second position of codons do not exist as they always cause an amino acid substitution or introduction of a stop codon. Table 1 lists all 190 noA sites: the gag ORF contains 28 noA sites, pol harbors 71 noA sites, env has 81 noA sites, and 10 noA positions are located in the nef ORF.
Figure 1. Clustering of noA sites in the HIV RNA genome. Overlapping windows of 30 nucleotides were defined to run with a step size of 1 nucleotide along the sequence counting noA sites in each window. A relatively high value for noA sites (4–6, Y-axis) at a particular nucleotide position (X-axis) indicates an enhanced proportion of noA nucleotides among its neighbors. The location in the sequence of ref-responsive element 34 is marked RRE, and an insert of frame (7520–7550) shows the spacing of individual noA sites. A cluster of as many as five noA sites is observed at the very core of the RRE structure near position 7542. SA marks the position of the splice acceptor sites SA7, SA7a, and SA7b SS. JPRT is the predicted RNA structure at the protease/RT junction in pol 18. The HIV genome map is shown at the top with the four overlap-free segments in the gag, pol, env, and nef ORFs that were probed for the occurrence of noA sites.

sites are present in the smaller nef ORF. Fig. 1 marks the HIV genome regions that were analyzed and shows a reasonable distribution of the noA sites, but a cluster near the end of the env ORF is immediately visible, which overlaps the structured RRE motif that is essential for HIV replication. The actual nucleotides at these 190 noA sites in the HIV genome are 86×U (45.3 per cent), 68×C (35.8 per cent), and 36×G (18.9 per cent). One might have expected more G if the RNA structure is a critical determinant of noA sites because G is part of the most stable G–C basepair and additionally can form the G–U pair. However, other sequence characteristics may explain the relatively infrequent G use at noA sites. For instance, G-suppression may relate to ‘CpG suppression’ that is common for genomes of many small eukaryotic viruses (Karlin, Doerfler, and Cardon 1994). Indeed, the second position of noA-carrying codons displays a relatively high proportion of the C nucleotide (12 As, 76 Cs, 17 Gs, and 85 Us), and hence, the possibility of CpG suppression is supported. CpG suppression has also been recognized in the HIV RNA genome, thereby shaping the codon usage (van Hemert and Berkhout 1995). With respect to a possible amino acid bias at the noA sites, all amino acids that have a noA-carrying codon are represented among these 190 noA codons, except Glu (GAA and GAG). We also determined the A frequency spectrum of all synonymous positions in the four ORFs. The frequency distribution among the categories 0 to 1 revealed that noA sites belong to a distinct codon population with a very low proportion of the A nucleotide, instead of being at the end of a gradually decreasing chain of A-nucleotide proportion (Fig. 2, the categories 0 and 0.1).

A possible raison d’être for the occurrence of noA sites may be the presence of a critically important sequence motif that plays a role during virus replication, in either the viral RNA genome that we analyzed or the double-stranded DNA copy that is generated by reverse transcription in HIV-infected cells. The identified noA sites were analyzed at three levels. First, we inspected whether clustering of multiple noA sites occurred, which may reveal the presence of an extended sequence motif or RNA structure that supports HIV replication. Second, we compared the base pairing potential of these rare noA sites versus the majority of regular
third codon positions in the detailed HIV RNA secondary structure models that have been determined by SHAPE technology (Watts et al. 2009). Third, we analyzed in silico the impact of forced mutagenesis of individual noA bases into the ‘forbidden’ A nucleotide on the local HIV RNA structure as predicted by the MFold tool.

The tendency for clustering of noA positions was investigated by monitoring the number of noA sites in 30-nucleotide windows along the viral RNA genome with a step size of a single nucleotide (Fig. 1). Regions with an enhanced proportion of noA sites (4–6 positions clustered) were detected around position 2095, 3326, and 4350 in pol, 7074, 7400, and 7862 in env, and position 8600 in the nef gene, but were absent in the gag ORF. Some of these clusters can already tentatively be associated with known viral RNA functions. The most prominent signal is a cluster of 36 noA sites around HIV genome position 7400 in the env ORF (see the inset in Fig. 1). This region is well-known as the essential Rev-Responsive Element or RRE, a highly structured RNA element that facilitates the nuclear export of unspliced and partially spliced HIV transcripts by interaction with the HIV Rev protein (Rausch, Le Grice, and Rev 2015). A bit further downstream starting at position 7862 we observed a cluster of 9 noA sites that coincide with the three frequently used splice acceptor sites SA7, SA7a, and SA7b at positions 7881, 7887, and 7915 (Purcell and Martin 1993). These correlations do cautiously support our hypothesis that the survey of noA sites can reveal important sequence and/or structure motifs. The HIV genome regions with an enhanced proportion of noA sites around positions 3326, 4350, 7074, and 8600 may thus represent yet unidentified HIV motifs of important RNA sequence or structure, which will be the topic of future studies. For instance, a noA cluster that overlaps with a conserved RNA structure termed JPRT of yet unknown function was identified at position 2095 in the pol ORF (Wang et al. 2008).

We next interrogated a possible interplay of the special noA sites and the local RNA structure requirements. To do so, we divided the C/G/U containing third base positions in two groups according to the presence (190) or absence (1119) of the noA label. For both groups, we plotted the base pairing probability based on the experimentally validated structure model for the HIV RNA genome (Fig. 3). Pairing probability values of individual positions were derived from the SHAPE-directed structure prediction of HIV RNA (Watts et al. 2009). These values range from 0 (unpaired) to 1 (paired) with steps of 0.1. The distribution histogram of Fig. 3 shows that the majority of pairing probability values of the 190 formal noA sites are >0.5, with a peak representation at 1.0 and a mean value of 0.57. In contrast, the 1119 regular sites without the noA label display a preference for values <0.5 with a peak representation at 0.1 and a mean value of 0.40. Thus, genome positions that do not allow the A nucleotide (noA sites) have a significantly higher base pairing probability than positions that allow the A nucleotide, which is consistent with the RNA structure hypothesis that states that the A nucleotide may be forbidden at specific positions as it would otherwise interfere with an important RNA structure.

These initial results suggest that noA sites may be prominently involved in the correct folding of the viral RNA genome. To further test this idea, we investigated the impact of in silico noA-to-A substitution on the local RNA structure for these three noA clusters, starting with the one at the essential RRE motif around position 7450. Introduction of A nucleotides at all 36 noA sites diminishes the folding free energy of this 300-nucleotide RNA sequence from −107.4 to −55.8 kcal/mole (Fig. 4), which indicates a prominent role of the noA sites in supporting this RNA secondary structure motif. When inspected in more detail, we scored a neutral effect on the RNA structure for seven A introductions, a stabilizing effect on the RNA structure for another seven A mutations and an RNA structure destabilizing effect for 22 positions. All individual stem segments of the RRE are affected by A introduction at noA sites, but the impact differs. For instance, the top domain II is heavily affected by seven A introductions at noA sites, which causes the loss of C–G pairs (4×), conversion of a strong C–G pair into a weaker A–U basepair (2×), and the loss of an A–U basepair (1×). In toto, the noA-to-A mutations have a dramatic effect on the folding of the RRE structure, which supports the notion that at least some of these noA sites are crucial to maintain the RRE motif in the proper conformation.

Splicing of HIV RNA is characterized by alternative splice sites and the expression of the many mRNA forms during the infection process that is regulated by the viral Rev protein (Purcell and Martin 1993). The three splice sites SA7, SA7a, and SA7b harbor several sequence-specific motifs for the binding of splicing factors and coincide with a cluster of 9 noA nucleotides (Fig. 1). Thus, sequence requirements may a priori form a better explanation
Figure 4. Predicted RRE structure in wild-type HIV RNA and upon introduction of ‘forbidden’ A nucleotides at noA sites. The RRE at position 7296–7596 encompasses the most prominent noA cluster in HIV RNA (Fig. 1). We manually exchanged the 36 noA bases (indicated by small arrows) for the forbidden A nucleotide, turning wild-type into the A mutant. Sequences were submitted to the MFold server for RNA secondary structure prediction. Major local structure alterations occur in the regions marked by Roman numbers I–IV. The thermodynamic stability (ΔG) is indicated, showing a 51.6 kcal/mole destabilization upon mutation of noA sites into A.

for the presence of this cluster of noA sites than RNA structure requirements. Anyhow, noA-to-A substitution was performed on a 105-nucleotide RNA fragment (7810–7914), followed by analysis of the predicted RNA structure by MFold (Fig. 5). The free energy of RNA folding for this RNA motif was affected by only 2.7 kcal/mole by the forced noA-to-A mutations. Close inspection of the structures revealed many differences, e.g. destabilization of the stem segment around SA7 and formation of many alternative basepairs in the central basepaired region. The 3′-terminal G7914 nucleotide of this region marks the start of the overlap between the env and tat/rev genes. In fact, the next nucleotide is the actual SA7b splice acceptor site and it is thus likely that G7914 is needed as part of a sequence motif for proper splicing. Consistent with this idea, the HIV sequence alignments indicated that nucleotides other than G are not allowed at this position.

A third cluster of noA sites maps at the RNA structure JPRT (Wang et al. 2008). Free energy values of the RNA structure predictions for wild-type and the A mutant differ by 15.4 kcal/mole (Fig. 6). Loop I is larger in the wild-type structure than for the A mutant, and the reverse is true for loop II. The stem structures III appeared quite similar, but all four noA-to-A substitutions introduced mismatches in the A mutant compared to the regular basepairs (3 A-U pairs and 1 U-A pair) in the wild-type structure. These data support a structural role for this RNA motif, which could be probed in future experiments. For a validation by means of mutagenesis, one could propose to test the same noA-to-A substitutions, which may help to reveal the unknown function of this structured RNA element JPRT in the HIV replication cycle.

The three noA clusters that we analyzed suggest that noA sites are intimately involved in the maintenance of important RNA sequences or structures, presumably to support HIV replication. It is likely that similar arguments hold for other noA clusters in the viral RNA, including novel unidentified structured HIV RNA motifs. We therefore looked into noA clusters elsewhere in the overlap-free regions of the pol, env, and nef genes. MFold predictions of wild-type and A-mutant structures were compared for the pol regions 1996–2176, 3306–3376, and 4286–4516, the env segment 7000–7200, and nef segment 8546–8746 (data not shown). Strikingly, introduction of A at these noA sites does invariably trigger a local destabilization of the respective RNA structure, which is in line with the proposed structural role of noA sites. Future research may be directed at dissecting the precise function of these candidate structured RNA signals in the HIV replication cycle.

Discussion

This study focused on silent codon positions in the HIV RNA genome that never or only rarely become A (termed noA sites), despite the pronounced evolutionary pressure to incorporate as many A nucleotides as possible in this viral RNA. We reasoned that such ‘resistance to A pressure’ likely indicates an evolutionary counterpressure to maintain important sequence information, as either a primary sequence element or a RNA structure motif that plays a role during virus replication. Although we will talk mostly about functional motifs in the viral RNA genome, we realize that retroviruses like HIV also use the DNA form and consequently such sequence motifs may also play a role as part of the viral DNA genome.

This in silico screen for noA sites in HIV RNA revealed clusters of noA sites in several known sequence and structure motifs, i.e. around the three SA7 splice sites and the well-known RRE structure that binds the Rev viral protein to facilitate nuclear export of unspliced and partially spliced HIV transcripts (Rausch, Le Grice, and Rev 2015). These positive hits do support the novel concept that one can identify important sequence/structure motifs by focusing on areas with noA sites. We revealed several new noA clusters in the HIV RNA genome (Fig. 1) that thus form candidate...
Figure 5. HIV RNA structure prediction of the splice acceptor region for wild-type and the A mutant. The region 7810–7914 encompassing splice acceptors SA7, SA7a, and SA7b constitutes the second prominent noA cluster in the HIV genome (Fig. 1). The splice acceptor sites are marked by black arrows, and the noA sites by red arrows. We manually exchanged the nine noA positions into the ‘forbidden’ A nucleotide in the A mutant. Sequences were submitted to the MFold server for RNA secondary structure prediction. The RNA structure was destabilized in the A mutant by only 2.7 kcal/mole, suggesting that RNA sequence motifs may be more decisive in this case.

Figure 6. RNA structure prediction of the wild-type and noA-mutated JPR7T motif. The protease/RT junction at 2013–2126 constitutes a cluster of ten noA sites (Fig. 1) for which the structure was recently proposed. RNA structure predictions are shown for wild-type RNA and the A mutant in which all ten noA positions (marked by red arrows) were converted into A (wild-type and A mutant). Typical differences between the structures are marked by Roman numbers I–III. The combined noA-to-A mutations affect the thermodynamic stability by 15.4 kcal/mole.

RNA/DNA signals to play important roles in HIV replication. Further experimental studies are needed to confirm these predictions. This should include a simple mutational analysis where one could consider the introduction of ‘forbidden’ A nucleotides at these silent noA positions exactly as in our in silico mutational strategy. As this will not affect the encoded protein sequence, one will probe
the functional contribution of the candidate noA position, either as a RNA/DNA sequence or possibly as a RNA structure motif. Such a mutational approach could hit a single noA residue or multiple positions in a noA cluster. The replication competence of such mutant viruses could first be tested in T cell lines, but one could also consider tests in primary T cells as some phenotypes are not apparent in transformed T cell lines. When a virus replication defect is scored, follow-up analyses are required to reveal the virus replication step that is affected by the novel RNA/DNA motif, e.g. in assays with subgenomic HIV constructs that quantitate a particular step of the virus replication cycle.

A similar in silico noA screen can be proposed for other RNA viruses, in particular those with a biased nucleotide composition of the genome and for which many isolates have been sequenced to fuel the analysis of potential noA sites. In fact, many viruses exhibit a biased nucleotide composition of their genome (Fros et al. 2017; van Hemert, van der Kuyl, and Berkhout 2016; van Hemert and Berkhout 2016; Berkhout and van Hemert 2015; Mochizuki et al. 2018; Ohara, and Roossinck; York et al. 2016; Selisko et al. 2018; Gaunt and Digard 2021). Although we focused here on a retrovirus with an RNA genome, the same principles relate to DNA viruses, where this in silico analysis could reveal not only critical DNA sequence motifs but also RNA signals that could play a role in virus biology as part of the viral transcripts. The requirement to have access to many viral genome sequences is usually met for viral pathogens that pose a significant health threat. For instance, some 32,800 complete SARS-CoV-2 genome sequences were deposited in GenBank within a year of the start of the pandemic, with an average of 90 full-length genome sequences (30,000 nucleotides per genome) per day.

The current noA screening method has some obvious limitations. We excluded regions of gene overlap as we want to focus on purely synonymous codon positions, a situation that is often blurred in regions of gene overlap. This may especially hamper the analysis of viruses as gene overlap occurs frequently in these condensed viral genomes. A second restriction concerns the fact that RNA viruses tend to cluster replication signals near the ends of the viral RNA genome to facilitate the correct copying of the full-length genetic information, more specifically the priming of minus- and plus-strand RNA synthesis. These terminal genome regions cannot easily be inspected for noA sites as they usually do not encode proteins, which thus forms a major limitation of this noA strategy to reveal novel viral replication signals.

For instance, HIV encodes many important replication signals in the 335-nucleotide-long 5′-untranslated region of the HIV RNA genome: the TAR hairpin that facilitates binding of the Tat protein to stimulate HIV transcription, the primer-binding site that facilitates binding of the tRNA primer for reverse transcription, and at least part of the packaging signal that controls the selective encapsidation of the viral RNA in virion particles (Keane and Summers 2016; Damgaard, Dyhr-Mikkelsen, and Kjems 1998, Huthoff and Berkhout 2001; Gallego et al. 2003; Keane et al. 2015). But this noA search seems ideally suited to screen for replication signals that are positioned more internally in the viral RNA genome.

We previously demonstrated profound accumulation of A nucleotides in the HIV RNA genome. In fact, not just silent codon changes underlie this A saturation as even non-silent codon changes are frequently used to raise the genomic A content. As a consequence, this extreme non-silent codon bias even results in a particular amino acid usage that reflects this A pressure on the HIV codons (Berkhout and van Hemert 1994; van Hemert and Berkhout 1995). We also revealed that this A pressure is more prominent in the unpaired regions of the HIV RNA structure in which more than 50 per cent A content can be reached (van Hemert, van der Kuyl, and Berkhout 2013). Surprisingly, mutagenesis toward a substantial increase (40.2 to 46.9 per cent) or decrease (31.7 to 26.3 per cent) of the A count in a 498-nucleotide polymerase gene segment of HIV-1 RNA did not affect the viral replication fitness in vitro, possibly because in vivo conditions are required to score this phenotype (Klaver et al. 2017). Accumulation of the A nucleotide in the HIV genome occurred over evolutionary times and may be caused by ApoBec3G-mediated cytidine deamination and subsequent G-to-A hypermutation (Malim and Bieniasz 2012) or may reflect the evolutionary pressure executed by other cellular restriction factors (Boso and Kozak 2020). The present study concerns genome positions where synonymous substitution into A could occur without an impact on the encoded protein, but for which such changes are not or only rarely observed. We revealed some 190 conserved noA sites in the 9 kb HIV RNA genome. These noA sites do display an enhanced base pairing probability in the established RNA secondary structure models of the HIV genome (Fig. 3), suggesting that RNA structure requirements may be a main driving force behind this noA phenomenon. This could not only reflect a general tendency of viruses to improve the stability of their viral RNA genome but also offer the opportunity for these viruses to build specialized functions based on RNA structure motifs. Apparently, two opposing evolutionary forces shape the HIV RNA genome.

First, there is a constant A pressure, which will ultimately affect the ability to form stable RNA structures and to encode active proteins. In contrast, specific noA sites survive despite this overwhelming A pressure by contributing to the presentation of important RNA sequence motifs and structures. HIV evolution should maneuver in sequence space and cope with both pressures, resulting in an A-enriched genome with local noA clusters that fulfill important replication functions as an RNA structure or sequence.

**Conflict of interest:** None declared.

**References**

Berkhout, B. and van Hemert, F. (2015) ‘On the Biased Nucleotide Composition of the Human Coronavirus RNA Genome’, *Virus Research*, 202: 41–7.

Berkhout, B. and van Hemert, F. J. V. (1994) ‘The Unusual Nucleotide Content of the HIV RNA Genome Results in a Biased Amino Acid Composition of HIV Proteins’, *Nucleic Acids Research*, 22: 1705–11.

Boso, G. and Kozak, C. A. (2020) ‘Retroviral Restriction Factors and Their Viral Targets: Restriction Strategies and Evolutionary Adaptations’, *Microorganisms*, 8: 1965.

Damgaard, C. K., Dyhr-Mikkelsen, H., and Kjems, J. (1998) ‘Mapping the RNA Binding Sites for Human Immunodeficiency Virus Type-1 Gag and NC Proteins within the Complete HIV-1 and -2 Untranslated Leader Regions’, *Nucleic Acids Research*, 26: 3667–76.

Ficarelli, M. et al. (2020) ‘CpG Dinucleotides Inhibit HIV-1 Replication through Zinc Finger Antiviral Protein (Zap)-dependent and -independent Mechanisms’, *Journal of Virology*, 94.

Foley, B. et al. (2013) *HIV Sequence Compendium 2013*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory.

Fros, J. J. et al. (2017) ‘CpG and UpA Dinucleotides in Both Coding and Non-coding Regions of Echovirus 7 Inhibit Replication Initiation Post-entry’, *Elife*, 6.

Gallego, J. et al. (2003) ‘Rev Binds Specifically to a Purine Loop in the SL1 Region of the HIV-1 Leader RNA’, *Journal of Biological Chemistry*, 278: 40385–91.
Gaunt, E. R., and Digard, P. (2021) ‘Compositional Biases in RNA Viruses: Causes, Consequences and Applications’, Wiley Interdisciplinary Reviews RNA, e1679.

Hall, T. A. (1999) ‘BioEdit: A User-friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT’, Nucleic Acids Symposium Series, 41: 95–8.

Huthoff, H. and Berkhout, B. (2001) ‘Two Alternating Structures of the HIV-1 Leader RNA’, RNA, 7: 143–57.

Karlin, S., Doerfler, W., and Cardon, L. R. (1994) ‘Why is CpG Suppressed in the Genomes of Virtually All Small Eukaryotic Viruses but Not in Those of Large Eukaryotic Viruses?’, Journal of Virology, 68: 2889–97.

Keane, S. C. et al. (2015) ‘RNA Structure. Structure of the HIV-1 RNA Packaging Signal’, Science, 348: 917–21.

Keane, S. C. and Summers, M. F. (2016) ‘NMR Studies of the Structure and Function of the HIV-1 5’-Leader’, Viruses, 8.

Klaver, B. et al. (2017) ‘Tolerates Changes in A-count in a Small Segment of the Pol Gene’, Retrovirology, 14: 43.

Kumar, S., Stecher, G., and Tamura, K. (2016) ‘MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 For Bigger Datasets’, Molecular Biology and Evolution, 33: 1870–4.

Malim, M. H. and Bieniasz, P. D. (2012) ‘HIV Restriction Factors and Mechanisms of Evasion’, Cold Spring Harbor Perspectives in Medicine, 2: a006940.

Mochizuki, T., Ohara, R., and Roossinck, M. J. (2018) ‘Large-Scale Synonymous Substitutions in Cucumber Mosaic Virus RNA 3 Facilitate Amino Acid Mutations in the Coat Protein’, Journal of Virology, 92.

Purcell, D. F and Martin, M. A. (1993) ‘Alternative Splicing of Human Immunodeficiency Virus Type 1 mRNA Modulates Viral Protein Expression, Replication, and Infectivity’, Journal of Virology, 67: 6365–78.

Rausch, J. W., Le Grice, S. F., and Rev, H. I. V. (2015) ‘Assembly on the Rev Response Element (RRE): A Structural Perspective’, Viruses, 7: 3053–75.

Selisko, B. et al. (2018) ‘Structural and Functional Basis of the Fidelity of Nucleotide Selection by Flavivirus RNA-Dependent RNA Polymerases’, Viruses, 10.

Takata, M. A. et al. (2017) ‘CG Dinucleotide Suppression Enables Antiviral Defence Targeting Non-self RNA’, Nature, 550: 124–7.

Vabret, N et al. (2014) ‘Large-Scale Nucleotide Optimization of Simian Immunodeficiency Virus Reduces Its Capacity to Stimulate Type I Interferon in Vitro’, Journal of Virology, 88: 4161–72.

Vabret, N., Bhardwaj, N., and Greenbaum, B. D. (2017) ‘Sequence-Specific Sensing of Nucleic Acids’, Trends in Immunology, 38: 53–65.

van der Kuyl, A. C. and Berkhout, B. (2012) ‘The Biased Nucleotide Composition of the HIV Genome: A Constant Factor in a Highly Variable Virus’, Retrovirology, 9: 92.

van Hemert, F. and Berkhout, B. (2016) ‘Nucleotide Composition of the Zika Virus RNA Genome and Its Codon Usage’, Virology Journal, 13: 95.

van Hemert, F., van der Kuyl, A. C., and Berkhout, B. (2016) ‘Impact of the Biased Nucleotide Composition of Viral RNA Genomes on RNA Structure and Codon Usage’, Journal of General Virology, 97: 2608–19.

van Hemert, F. J. and Berkhout, B. (1995) ‘The Tendency of Lentiviral Open Reading Frames to Become A-rich: Constraints Imposed by Viral Genome Organization and Cellular tRNA Availability’, Journal of Molecular Evolution, 41: 132–40.

van Hemert, F. J., van der Kuyl, A. C., and Berkhout, B. B. (2013) ‘The A-nucleotide Preference of HIV-1 in the Context of Its Structured RNA Genome’, RNA Biology, 10: 211–5.

Wang, Q. et al. (2008) ‘Evidence of a Novel RNA Secondary Structure in the Coding Region of HIV-1 Pol Gene’, RNA, 14: 2478–88.

Watts, J. M. et al. (2009) ‘Architecture and Secondary Structure of an Entire HIV-1 RNA Genome’, Nature, 460: 711–6.

York, A. et al. (2016) ‘The RNA Binding Specificity of Human APOBEC3 Proteins Resembles that of HIV-1 Nucleocapsid’, PLoS Pathogens, 12: e1005833.

Zuker, M. (2003) ‘Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction’, Nucleic Acids Research, 31: 3406–15.