Virus Research xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

# On the nucleotide composition and structure of retroviral RNA genomes

Formijn van Hemert, Antoinette C. van der Kuyl, Ben Berkhout*

*Laboratory of Experimental Virology, Department of Medical Microbiology, Center for Infection and Immunity Amsterdam (CINIMA), Academic Medical Center, University of Amsterdam, The Netherlands*

ARTICLE INFO

*Article history:*
Available online xxx

*Keywords:*
Retrovirus
RNA genome
Nucleotide composition
RNA structure
Virus evolution

ABSTRACT

Retroviral RNA genomes display a rich variety in their nucleotide composition. For instance, the single-stranded RNA genome of human T cell leukemia virus (HTLV-1) is C-rich and G-poor and that of the human immunodeficiency virus (HIV-1) is A-rich and C-poor. Animal retroviruses add further variation to this unexplained, but many times remarkable virus-specific property. We previously described that the nucleotide bias is even more extreme in the unpaired regions of the structured HIV-1 RNA genome, which has been probed by SHAPE technology. We now document that the same trend is apparent for the MFold-predicted RNA structure of HIV-1 RNA and subsequently investigated the predicted structures of the RNA genomes of other retroviruses. We conclude that all virus-specific signatures are enhanced for the unpaired nucleotides in the RNA genome. Consequently, the differences in nucleotide count between the diverse human and animal retroviruses are further exposed in the single stranded genome regions. We used a skew analysis to visualize these striking differences in nucleotide usage. Evolutionary events responsible for these nucleotide signatures will be discussed.



## 1. Introduction

We and others documented a tendency of the HIV-1 RNA genome to have a biased nucleotide composition with more than 36% A and only 17% C (Berkhout and van Hemert, 1994; van Hemert and Berkhout, 1995; Kypr and Mrazek, 1987; Kypr et al., 1989; Zoubak et al., 1992; Bronson and Anderson, 1994; Zsiros et al., 1999). Although the initial suggestion was that HIV-1 uses a special set of codons to translate its proteins (Kypr and Mrazek, 1987), this was subsequently demonstrated to reflect a more general bias in favor of the A nucleotide (van Hemert and Berkhout, 1995). In fact, this trend even influences the amino acid usage as HIV-1 prefers A-rich codons, sometimes irrespective of the encoded amino acids (Berkhout and van Hemert, 1994). The human Schlafen protein SLFN11 was recently proposed to have antiviral activity by selectively blocking viral protein synthesis in HIV-infected cells by means of codon-bias discrimination (Li et al., 2012). All members of the lentivirus family, including viruses of human and animal origin, share this typical nucleotide bias, which is also highly conserved in evolution (van der Kuyl and Berkhout, 2012). Other members of the Retroviridae family have distinct, but also highly biased nucleotide signatures (Berkhout et al., 2002). The specifics are listed in Table 1 for representative members of the different retrovirus groups.

The mechanisms responsible for these biases currently remain unknown. The bias could be caused by a preferred mutational spectrum of the enzymes that copy the retroviral genome, that is the viral reverse transcriptase and the cellular RNA polymerase II, or enzymes that can perform editing events (e.g. Apobec or ADAR enzymes) (Vartanian et al., 1994; Meyerhans et al., 1994; Martinez et al., 1994; Bishop et al., 2004; Deforche et al., 2007; Wood et al., 2009; Berkhout and de Ronde, 2004; Belanger et al., 2014). Alternatively, retroviruses may have evolved into a specific corner of sequence space by selective pressure. For instance, a biased RNA genome may present a molecular signature that is either recognized by the virus to selectively package its genome in virion particles or that is able to modulate cellular antiviral responses such as the interferon response (Vabret et al., 2012; van der Kuyl and Berkhout, 2012; Fenton-May et al., 2013; van Bel et al., 2014).

We recently took this analysis of the nucleotide composition one step further by inspection in the context of the RNA secondary structure. This is possible for the HIV-1 RNA genome because its structure was determined by the SHAPE technology (Watts et al., 2009). We reported excessive A-accumulation in the unpaired domains of the folded RNA molecule, contrasting with the absence of such bias in the base-paired sequences (van Hemert et al., 2013).

* Corresponding author. Tel.: +31 205664822.
*E-mail addresses:* b.berkhout@amc.uva.nl, commoffice@amc.uva.nl (B. Berkhout).

http://dx.doi.org/10.1016/j.virusres.2014.03.019
0168-1702/

**Table 1**
Genome signature of representative retroviruses.

| Genus | Species | Genbank ID | Nucleotide signature |
|---|---|---|---|
| Alpharetrovirus | Rous sarcoma virus (RSV) | J02342 | Moderately biased[a] |
| Betaretrovirus | Mason-Pfizer monkey virus (MPMV) | M12349 | High A, low G |
| | Mouse mammary tumor virus (MMTV) | NC_001503 | Moderately biased[a] |
| Gammaretrovirus | Moloney murine leukemia virus (MoMuLV) | J02255 | Moderately biased[a] |
| Deltaretrovirus | Human T-lymphotropic virus 1 (HTLV-1) | NC_001436 | High C, low G |
| | Human T-lymphotropic virus 2 (HTLV-2) | M10060 | High C, low G |
| Epsilonretrovirus | Walleye dermal sarcoma virus (WDSV) | NC_001867 | High A, low G |
| Lentivirus | Human immunodeficiency virus 1 (HIV-1) | M19921 | High A, low C |
| | Simian immunodeficiency virus (SIV) | M33262 | High A, low C |
| Spumavirus | Human foamy virus (HFV) | Y07725 | High A/U, low C/G |

[a] Moderately biased: each nucleotide percentage between 21 and 30%.

In this study, we set out to perform a similar analysis for the diverse retroviral species that exhibit a strikingly different nucleotide composition. A problem is the absence of experimentally probed RNA structure models for these retroviruses. We therefore verified whether the notable results described for the experimentally probed HIV-1 RNA structure could be reproduced with a computer-predicted structure of this viral RNA genome. Next, we set out to predict the structure of the retroviral RNA genomes listed in Table 1 and subsequently analyzed the nucleotide signature in the single-stranded (ss) and double-stranded (ds) domains.

## 2. Materials and methods

Nucleotide sequences of the prototype retroviruses listed in Table 1 were taken from Genbank. We analyzed the retroviral RNA genomes, which are a bit smaller than the respective DNA genomes. MFold (Zuker, 2003) and incidentally RNAstructure (Reuter and Mathews, 2010) were used with default settings for RNA secondary structure prediction. SHAPE-mediated HIV-1 and SIVmac239 structure data were available from Watts et al. (Watts et al., 2009) and Pollom et al. (Pollom et al., 2013), respectively. The ss-count file of a MFold output supplied the number of folded structures (50 maximally) including a frequency value for each individual nucleotide of being unpaired in this collection of structures. We scored a nucleotide as unpaired (single-stranded, "ss") if half or more than half of the structure models reported its position as "ss". Nucleotides with a ss-count value below this criterion were scored as being paired (double-stranded, "ds"). Discrimination between ss and ds nucleotides was done in Excel and a fasta file was created. Nucleotide composition was determined by means of MEGA 5.2 (Tamura et al., 2011).

The size limit for submission to the MFold server is 9000 nucleotides. In case a viral genome length exceeded this limit, a bipartition was made with an overlap of 1000–2000 nucleotides between positions 7000 and 9000. The ss-count data of both submission output files were arithmetically averaged at the region of overlap before the ss/ds discrimination was performed. We used the MFold ct file of the top 1 structure model to assess basepair identity in retroviral RNA (regular Watson–Crick and G–U/U–G pairs). Partial sequence files were reconstituted at a site near the center of the overlap to minimize folding artifacts near the borders of the submitted sequences.

Base composition analysis along the RNA genome length and the accompanying ss and ds fasta files was performed by the method of cumulative skew diagrams in overlapping windows (Grigoriev, 1998). For normalization purposes, windows were defined around 1% of the sequence length with a step size of 20% of the window size. The SKEW software does not allow for windows exceeding 100 nucleotides and as a result, the "all nts" skew values of WDSV and HFV with genome lengths of 12.709 and 13.243 nucleotides, respectively, may be slightly overestimated. Calculations were done in Excel.

## 3. Results

### 3.1. The nucleotide distribution in predicted versus experimentally probed HIV-1 RNA structure models

Whereas our previous analysis focused on the HIV-1 RNA genome, we now present a more comprehensive analysis of other members of the Retroviridae family (Table 1). Prototype representatives were selected for the alpharetrovirus, betaretrovirus, gammaretrovirus, deltaretrovirus, epsilonretrovirus, lentivirus and spumavirus groups. This set includes several retroviruses that prefer the A or C nucleotide and two retroviruses with a moderately biased nucleotide composition (RSV and MoMuLV). For HIV and HTLV, we included two subgroups that show considerable similarity in a phylogenetic analysis (HIV-1 and SIVmac239, HTLV-1 and HTLV-2, respectively).

As no experimentally probed RNA structure models are available for these viruses except for HIV-1 and SIV (Watts et al., 2009; Pollom et al., 2013), we first verified whether a computer-generated RNA structure is a useful substrate for such an analysis. To do so, we compared the ss/ds distribution for all four nucleotides in the structures obtained with the MFold (50 folds) and RNAstructure (20 folds) algorithms for RNA secondary structure prediction to that of the SHAPE-determined HIV-1 RNA fold (Pollom et al., 2013), which actually was built using the RNAstructure software. A good correlation is apparent for the distribution of ss and ds nucleotides among the prediction programs and the combined SHAPE + RNAstructure exercise (Table 2). This correlation holds for all nucleotides, including the A-nucleotide, which constitutes 53.7, 53.5 and 51.3% of the ss pool and only 20.4, 22.1 and 19.7% of the ds compartment of HIV-1 nucleotides according to MFold, RNAstructure and SHAPE + RNAstructure. These results reinforce the idea that predicted RNA structure models can be used in a first attempt to dissect structural features of the nucleotide distribution. In this study, we continued with the MFold-predicted structure of HIV-1 RNA and compared that to the genome of other retroviruses. The ss-count output file provides for each nucleotide a probability value of being unpaired in the top models (50, except for RSV (40) and HFV (45)) generated by a single MFold folding job (see Section 2).

### 3.2. The nucleotide composition drives the basepair usage in retroviral RNA genomes

We next probed whether the typical nucleotide compositions do influence the type of basepairs used in the structured RNA genomes. The previous HIV-1 study noticed a trend toward the more frequent usage of the more stable basepairs (GC + CG > AU + UA > GU + UG). This unequal basepair usage correlates with the G and C preference

Please cite this article in press as: van Hemert, F., et al., On the nucleotide composition and structure of retroviral RNA genomes. Virus Res. (2014), http://dx.doi.org/10.1016/j.virusres.2014.03.019