

Interplay between RNA Structure and Protein Evolution in HIV-1

Rafael Sanjuán^{*,1,2} and Antonio V. Bordería³

¹Institut Cavanilles de Biodiversitat i Biologia Evolutiva and Departamento de Genética, Universitat de València, Spain

²Centro Superior de Investigación en Salud Pública (CSISP), Area de Genómica y Salud, Spain

³Institut Pasteur, Department of Virology, Viral Populations and Pathogenesis Group. Paris cedex 15, France

*Corresponding author: E-mail: rafael.sanjuan@uv.es.

Associate editor: Daniel Falush

Abstract

The genomes of many RNA viruses contain abundant secondary structures that have been shown to be important for understanding the evolution of noncoding regions and synonymous sites. However, the consequences for protein evolution are less well understood. Recently, the secondary structure of the HIV-1 RNA genome has been experimentally determined. Using this information, here we show that RNA structure and proteins do not evolve independently. A negative correlation exists between the extent of base pairing in the genomic RNA and amino acid variability. Relaxed RNA structures may favor the accumulation of genetic variation in proteins and, conversely, sequence changes driven by positive selection at the protein level may disrupt existing RNA structures. We also find that breakage of RNA base pairs might impose a fitness cost to drug resistance mutations in the protease and reverse transcriptase genes, thereby limiting their spread among untreated patients. Characterizing the evolutionary trade-offs between the selective pressures acting at the RNA and protein levels will help us to better understand the variability and evolution of HIV-1.

Key words: RNA structure, RNA virus, molecular evolution, selective constraint, drug resistance, fitness trade-off.

Introduction

RNA structure plays an important role in the molecular biology of the cell. This is true for the large number of noncoding RNAs expressed in eukaryotes (Cawley et al. 2004; Voinnet 2009), but also for mRNA. In the latter, RNA structure controls the efficiency of translation by several mechanisms, including 5' UTR-mediated initiation and stabilization (Kudla et al. 2009), riboswitches (Tucker and Breaker 2005) or interactions with the double-stranded RNA-activated protein kinase (Kaempfer 2003), and can modulate the dynamics of protein folding by ribosome pausing (Watts et al. 2009) or even modify protein sequence by inducing ribosome slippage (Xu et al. 2001). RNA structures are particularly relevant to viruses with single-stranded RNA genomes (Frankel and Young 1998; Pelletier and Sonenberg 1988; Damgaard et al. 2004). For instance, the initiation of HIV-1 reverse transcription is mediated by binding of a tRNA to the primer binding site situated in a specific stem-loop structure of the 5' long terminal repeat, where other functionally important structures have been described (Damgaard et al. 2004). More generally, the structure of the viral genomic RNA has been associated with viral host-defense evasion and persistence (Simmonds et al. 2004; Tellam et al. 2008), replicative capacity, and cross-species transmission (Brower-Sinning et al. 2009).

As a consequence of selective pressures acting at the RNA level, nucleotide sites involved in forming intramolecular base pairs in the viral RNA tend to show less variability and lower rates of evolution than those not forming pairs,

as demonstrated for HIV-1 (Le et al. 1988; Le et al. 1989; Yoshida et al. 1997), Visna virus (Braun et al. 1987), hepatitis C virus (Contreras et al. 2002; Tuplin et al. 2002), GB virus (Simmonds and Smith 1999), influenza A virus (García et al. 1996), or the cucumber mosaic virus satellite RNA (Rodríguez-Alvarado and Roossinck 1997). Other salient features associated with RNA structure are the excess of transitions over transversions at sites forming base pairs, which has been interpreted in terms of compensatory evolution (Knies et al. 2008), and the increase in G+C content resulting from selection for thermostability (Schultes et al. 1997; Smit et al. 2009). However, these patterns have been established mainly for noncoding sequences or synonymous variation within coding regions, whereas the interplay between RNA structure and protein evolution is less well understood.

Recently, the secondary structure of the entire HIV-1 genome has been determined experimentally using high-throughput selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) reactivity (Watts et al. 2009). This method has allowed the identification of RNA base pairs with high confidence and offers a unique opportunity to study the implications of RNA structure for HIV-1 variability and evolution. Here, by combining these experimental data with molecular evolution analyses, we demonstrate that amino acid sites encoded by structure-forming RNA are consistently less variable across the genome and evolve more slowly than those for which such RNA-level selection is absent. This finding is further strengthened by the analysis of specific RNA motifs in

the genome, the patterns of variation in the HIV-1 surface protein, and the epidemiology of drug resistance. Overall, the data suggest a reciprocal interference between selection at the RNA and protein levels, in which the need to maintain certain RNA structures may curtail protein evolution and, conversely, the ability of the genomic RNA to form base pairs may be restricted by selection at the protein level.

Materials and Methods

SHAPE Reactivity

We used SHAPE reactivity values from sites 12 to 9142 of the NL4-3 reference HIV-1 genome (Genbank accession AF324493) available from Watts et al. (2009) as a measure of the probability of base pairing for each nucleotide. The RNA secondary structure model proposed by Watts et al. (2009) was not used here because it was inferred using co-variation evolutionary analysis (Pedersen et al. 2004), thus breaking the statistical independence between variability and RNA structure.

Predicted RNA Secondary Structures

Regions with anomalously low- or high-average SHAPE reactivity value were searched in the HIV-1 genome using sliding windows of various sizes (25, 50, 100, 200, and 300 nucleotides). Lowest free energy structures for each of these regions were predicted with the algorithm implemented in RNAstructure v5.1 (Reuter and Mathews 2010) (<http://rna.urmc.rochester.edu/RNAstructure.html>), using SHAPE reactivity values as pseudoenergy constraints (parameters were set to default values). The regions identified by the sliding window were cut or extended manually to accommodate predicted stems and loops. Structures were visualized using CLC RNA Workbench (CLCbio).

HIV-1 Alignments and Molecular Evolution

A hundred nucleotide sequences of HIV-1 subtype B were obtained from the Los Alamos HIV Sequence Database (<http://www.hiv.lanl.gov>) for each gene. The translated sequences were aligned with Muscle3.6 (Edgar 2004) (<http://www.drive5.com/muscle>), manually edited, and converted back to nucleotides. Sequences with premature stops codons were not considered. Nucleotide diversity was calculated from the final alignment using DAMBE (Xia and Xie 2001) (<http://dambe.bio.uottawa.ca/dambe.asp>). Estimation of synonymous (d_S) and nonsynonymous (d_N) substitution rates and categorization of codons as positively ($d_N > d_S$) or negatively ($d_N < d_S$) selected was done using the SLAC/GARD method (Kosakovsky Pond and Frost 2005) implemented in the HYPHY2.0 package (<http://www.datam0nk3y.org/hyphy>). An Excel spreadsheet containing site-by-site results is available upon request.

Drug Prevalence Data

Drug resistance prevalence for mutations in the protease and reverse transcriptase genes in either naive or treated patients were obtained from the HIV Drug Resistance Mutation Database of the Stanford University ([\[hivdb.stanford.edu\]\(http://hivdb.stanford.edu\)\). In total, we considered 337 mutations \(323 substitutions and 14 insertions or deletions\). Of the total 539 amino acid residues examined, 125 contained substitutions and, in 67 of these, there were two or more different mutations reported \(the maximum being 13 different mutations at residue 215 of the RT gene\). Using the NL4-3 sequence as a reference, we identified the simplest nucleotide mutational pathway leading to the observed amino acid substitution. In 242 cases, a single substitution at one of the three codon sites was sufficient. Double \(64 cases\) and triple \(6 cases\) nucleotide substitutions were also identified. In total, the 323 amino acid substitutions occurring at 125 residues were explained by 388 substitutions occurring at 209 nucleotide sites. Knowing the exact nucleotide involved in each drug resistance mutations allowed us to test the correlation between mutation prevalence and RNA structure more accurately than by relying on the amino acid sequence. However, due to the degeneracy of the genetic code, 45 of the total 388 nucleotide substitutions were ambiguous, that is, there were two or three possible changes at a given site leading to the same amino acid substitution. Overall prevalence among treated patients was directly taken from the above database. Prevalence among untreated patients was only available by viral genotype, so we obtained the overall prevalence as the average prevalence across genotypes, weighting by the number of cases of each. A spreadsheet containing drug resistance prevalence values for each nucleotide site is available upon request.](http://</p>
</div>
<div data-bbox=)

Results

General Association between RNA Structure and Protein Variability

The SHAPE reactivity of each site of the HIV-1 genome depends on the probability that the site is forming an intramolecular base pair. High SHAPE reactivity values (typically >0.5) indicate steric availability to chemical modification and thus that the site is probably not forming a base pair, whereas low reactivity values (<0.25) indicate the contrary (Watts et al. 2009). Using 100 full-length subtype B genomes, we first calculated the nucleotide diversity at each site of the *gag*, *pol*, *env*, *vif*, *vpr*, *vpr* and *nef* genes. Whereas substitutions at first- and third-codon positions can be either synonymous or nonsynonymous, all substitutions at second-codon positions are nonsynonymous and hence allow us to quantify protein sequence variation in a simple way. We found that SHAPE reactivity correlated positively with second codon–position diversity (Spearman's $\rho = 0.093$, $N = 2564$, $P < 0.001$). Consistently, polymorphic second-codon sites tended to have higher SHAPE reactivity (median 0.38) than invariant ones (median 0.31; Mann–Whitney test: $N = 2564$, $Z = -4.276$, $P < 0.001$).

To more precisely ascertain the effect of RNA structure on protein evolution, we estimated synonymous (d_S) and nonsynonymous (d_N) substitution rates in a codon-wise manner. SHAPE reactivity correlated positively with d_N across the whole genome region studied ($\rho = 0.109$, $N = 7688$, $P < 0.001$). The correlation held for *pol* ($\rho = 0.127$, $P <$

0.001), *vif* ($\rho = 0.174$, $P < 0.001$), *vpr* ($\rho = 0.169$, $P = 0.014$), *env* ($\rho = 0.161$, $P < 0.001$), and *nef* ($\rho = 0.122$, $P = 0.002$) genes individually was marginally significant for *gag* ($\rho = 0.046$, $P = 0.099$) and not significant for *vpu* ($\rho = -0.073$, $P = 0.360$). Furthermore, when we categorized each site as polymorphic ($d_N > 0$) or invariant ($d_N = 0$) at the protein level, we observed that polymorphic sites had a significantly higher SHAPE reactivity than invariant ones (medians 0.39 and 0.29, respectively; $N = 7688$, $Z = -10.446$, $P < 0.001$), and analysis of individual genes confirmed this pattern (fig. 1a). SHAPE reactivity also correlated positively with d_S across the genome, confirming previous findings (Le et al. 1988; Le et al. 1989; Yoshida et al. 1997), but the correlation was weaker ($\rho = 0.026$, $N = 7688$, $P = 0.024$) than the one between SHAPE reactivity and d_N and, on a per-gene basis, was significant only for *env* ($\rho = 0.073$, $P = 0.001$) and *nef* ($\rho = 0.112$, $P = 0.005$), and marginally so for *vpu* ($\rho = 0.154$, $P = 0.053$). Categorization of each site as polymorphic ($d_S > 0$) or invariant ($d_S = 0$) yielded a similarly weak association with SHAPE reactivity compared with the one observed for nonsynonymous sites (fig. 1b).

We then searched the genome for protein-coding regions with particularly high or low SHAPE reactivity values (fig. 2). The predicted secondary structure of the regions identified as having low reactivity contained abundant base pairs and tended to form stems, whereas the high-reactivity regions were predominantly looped out. The total median diversity for the nucleotide sites predicted to form base pairs in the low-reactivity regions was 0.008, whereas the unpaired sites in high-reactivity regions were significantly more diverse (median 0.055; Mann–Whitney test: $N = 492$, $Z = -6.595$, $P < 0.001$), confirming the association between RNA structure and variability. Furthermore, the high-reactivity regions showed both greater d_N ($Z = -7.475$, $P < 0.001$) and d_S ($Z = -5.636$, $P < 0.001$) values (table 1). It is also of notice that the regions identified were located preferentially in *pol*, *env*, and *nef*, and that these are the genes for which the association between RNA structure and diversity was found to be more significant (fig. 1). Additional analyses for two of these genes are presented below.

Interplay between RNA Structure and the Evolution of *env* and *pol* genes

The immune system exerts a strong pressure on the *env* gp120 surface envelope protein through neutralizing antibodies and the cytotoxic T-cell response (Wei et al. 2003; Kawashima et al. 2009), resulting in many positively selected codons. The gp120 protein is often divided into five hypervariable regions (V1 to V5) that alternate with five relatively conserved regions (C1 to C5). Confirming the importance of selective pressures acting at the protein level (immune selection or others), we found that 23.8% of codons belonging to V1–V5 regions were under significantly positive selection, whereas this percentage dropped to 12.2 for C1–C5 (Fisher's exact test: $P < 0.001$). However, this kind of selection does not appear to be the only factor contributing to V1–V5 hypervariability. We found that SHAPE reactivity was higher for the V1–V5 regions (median 0.46)

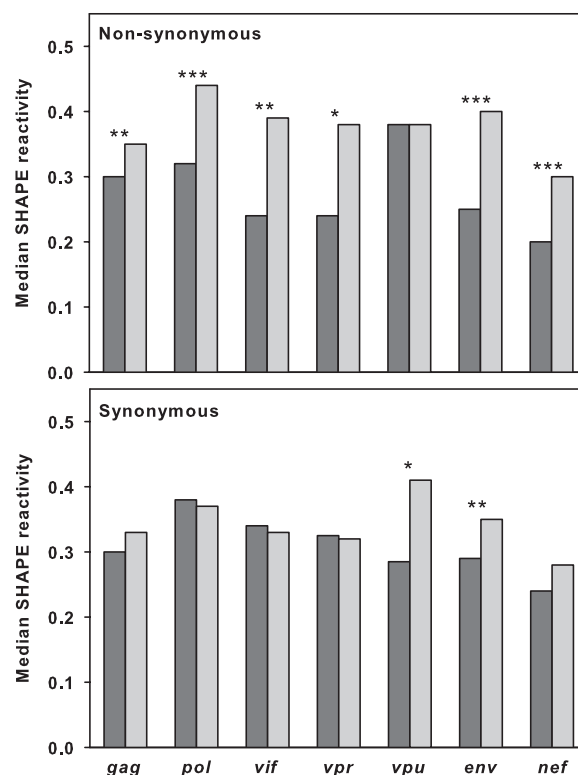


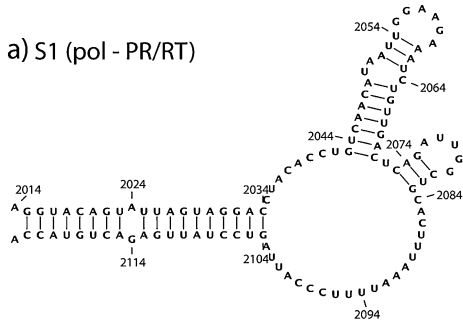
Fig. 1. Genome-wide association between RNA structure and variability. Upper panel: median SHAPE reactivity values for invariant ($d_N = 0$, dark gray) versus polymorphic ($d_N > 0$, light gray) amino acid sites. Lower panel: same graph for synonymous variation. Asterisks denote statistical significance (***Mann–Whitney test, $P < 0.001$; ** $P < 0.01$; * $P < 0.1$).

than for C1–C5 (median 0.35; $N = 1442$, $Z = -4.0976$, $P < 0.001$) and was particularly elevated at V1 and V2 (median 0.53). Furthermore, there was a positive correlation between d_N and SHAPE reactivity across the entire gp120 gene ($\rho = 0.164$, $N = 1442$, $P < 0.001$). Consistent with the hypothesis of selection acting on RNA structure, the V1–V5 regions also showed higher d_S values than C1–C5 ($N = 1442$, $Z = -9.925$, $P < 0.001$). Therefore, relaxation of the RNA structure appears to facilitate the accumulation of genetic variation in gp120, thus contributing to the existence of hypervariable regions.

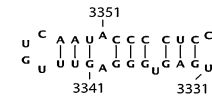
The *pol* gene is the target of antiretroviral drugs directed against the protease (PR) and the reverse transcriptase (RT). According to data from the Stanford HIV Drug Resistance Mutation Database, drug resistance mutations had a median prevalence of 2.9% per mutation among treated patients and a roughly 20 times lower prevalence among untreated patients (0.13%). Consistent with treated-to-naïve transmission of drug resistance (Hue et al. 2009), 87% of the mutations found in treated patients were also present in untreated patients and prevalence values in these two groups were positively correlated ($\rho = 0.541$, $N = 209$, $P < 0.001$). The spread of drug resistance among treated individuals is an obvious consequence of the strong selective pressure exerted on the PR and RT proteins. In contrast, their prevalence in the untreated population should be also influenced by the fitness cost of resistance in the absence of the drug, and this

STEMS

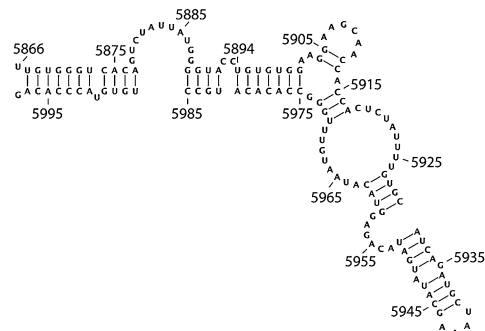
a) S1 (pol - PR/RT)



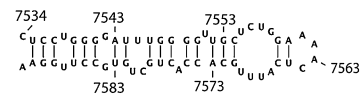
b) S2 (pol - RT)



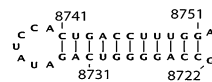
c) S3 (env - gp120)



d) S4 (env - gp41)

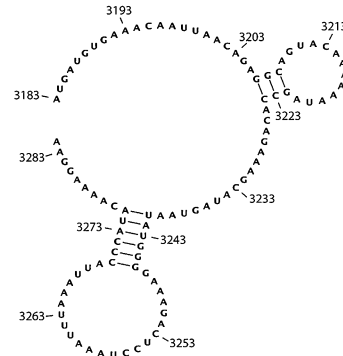


e) S5 (nef)

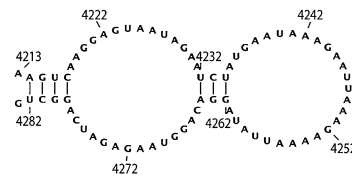


LOOPS

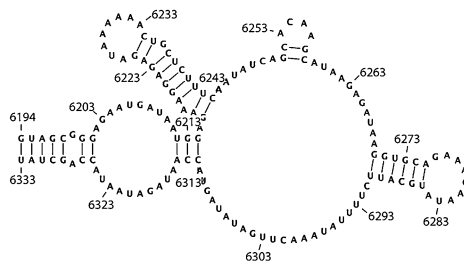
f) L1 (pol - RT)



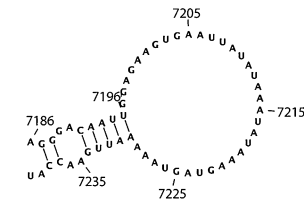
g) L2 (pol - IN)



h) L3 (env - gp120)



i) L4 (env - gp120)



j) L5 (nef)

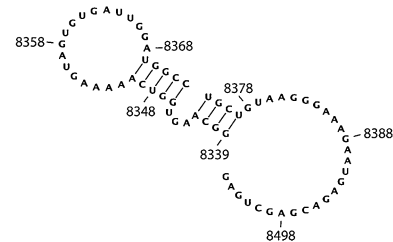


Fig. 2. Predicted RNA secondary structures for ten representative regions with low (stems) versus high (loops) SHAPE reactivity within protein-coding regions. The gene and nucleotide positions relative to the NL4-3 genome are shown for each structure. Several of the structures match with known motifs (Watts et al. 2009). S1: longest continuous helix; S3: *env* signal peptide stem; S4: part of the Rev-responsive element stem I; L2: central polypurine tract.

Table 1. Median SHAPE reactivity, nucleotide diversity, d_N and d_S of the ten RNA secondary structures shown in figure 2

Name	Start ^a	End ^a	SHAPE ^b	Diversity ^b	d_N^b	d_S^b
S1	2015	2122	0.080	0.010	0.00	2.00
S2	3332	3356	0.130	0.008	0.48	3.80
S3	5867	5997	0.100	0.008	0.00	3.00
S4	7534	7592	0.080	0.000	0.00	2.08
S5	8723	8751	0.045	0.024	1.10	4.26
L1	3183	3283	0.710	0.024	0.95	4.00
L2	4219	4278	0.640	0.016	0.61	2.61
L3	6202	6325	0.640	0.201	9.59	5.88
L4	7196	7229	0.785	0.016	0.00	3.26
L5	8343	8404	0.655	0.136	5.49	7.45

^a NL4-3 genome position.

^b For structures S1–S5, medians were calculated over all sites predicted to form base pairs, whereas for L1–L5, they were calculated over all sites predicted not to form pairs.

cost might be caused by the disruption of base pairs in the RNA structure of the viral genome, among other factors. Supporting this hypothesis, there was a weak but significantly positive correlation between SHAPE reactivity and the prevalence of drug resistance mutations among untreated patients ($\rho = 0.158$, $N = 183$, $P = 0.032$), and mutations within the highest prevalence quartile were located at sites with higher SHAPE reactivity values (median 0.39) than those within the lowest prevalence quartile (median 0.18; $N = 90$, $Z = -2.442$, $P = 0.015$). In contrast, as expected, no significant correlation was found between SHAPE reactivity and mutation prevalence among treated patients ($\rho = 0.037$, $N = 199$, $P = 0.599$).

Discussion

Our results suggest that the selective pressure for maintaining the genomic RNA structure constrains the variability and evolution of HIV-1 not only at synonymous sites, but also at the protein level. An alternative explanation for the association between SHAPE reactivity and genetic variability is that some of the nucleotide substitutions favored by protein-level selection may disrupt existing RNA base pairs. Some of the data seem compatible with this second interpretation. For instance, on a genome-wide basis, SHAPE reactivity correlated more strongly with d_N than with d_S , whereas RNA structure constraints should have a similar effect on these two quantities. On the other hand, the analysis of specific stem-loop structures indicated that both d_N and d_S showed a highly significant correlation with SHAPE reactivity. Furthermore, if selection at the protein level was the main factor explaining the association between protein variability and SHAPE reactivity, positively selected codons should show higher SHAPE reactivity than nonselected or negatively selected codons throughout the entire genome, but this was clearly not the case (medians 0.35 and 0.34, respectively; $N = 7013$, $Z = -0.366$, $P = 0.715$), except maybe for the gp120 protein where the V1–V5 regions are both under strong positive selection and show the highest SHAPE reactivity. Finally, the association between drug resistance mutation prevalence in untreated patients and SHAPE reactivity cannot be easily explained in terms of protein-level selection. Therefore,

the most likely scenario is that the causal link between RNA structure and protein evolution goes in both directions, such that the selective pressure for the maintenance of RNA base pairing restricts protein variability and, conversely, positive selection favoring amino acid change can result in the disruption of RNA base pairs.

An interesting possibility is that the formation of locally unstructured RNA might be a mechanism by which HIV-1 and other RNA viruses could modulate their variability (Braun et al. 1987; Le et al. 1988, 1989; Schinazi et al. 1994). In general, increasing the number of segregating polymorphisms allows for a more rapid response to selection, and a preadaptation of the virus to generate escape mutants in the *env* gene has been hypothesized (Le et al. 1988). It is also interesting to note that highly variable regions such as the gp120 V1–V5 are sometimes flanked by stable RNA helices that could be considered as structural insulators of variability (Watts et al. 2009). Although selection appears to be the main factor explaining the association between variability and RNA structure, it is also conceivable that RNA structure might modulate the mutation rate. Single-stranded nucleic acids are more prone to chemical damage (Lindahl and Nyberg 1974), strand cleavage (Parthasarathi et al. 1995), and to APOBEC3G-mediated edition (Yu et al. 2004), and a direct link between RNA structure and polymerase fidelity has been postulated (Ji et al. 1994; Yoshida et al. 1997; Contreras et al. 2002). Future work may elucidate whether replication is intrinsically more error prone in nonstructured RNAs.

Previous work has assessed the role played by non-coding viral RNA structures in the regulation and initiation of replication, transcription, and translation (Carter and Saunders 2007). Evidence that RNA structures can form elsewhere in the genome (Simmonds et al. 2004; Watts et al. 2009) and that they are important for determining not only synonymous variation but also protein evolution, as shown here for HIV-1, opens new research avenues. In future work, the observation that some viral genes or genome regions show higher levels of variation than others should not only be interpreted solely in terms of protein selection, but also in terms of selection at the RNA level. Furthermore, conflicts between these two levels of selection may constitute a previously unrecognized constraint to RNA virus evolvability. On the experimental side, the determination of genome-wide RNA structures for viruses other than HIV-1 will allow further testing of the results obtained here. On the bioinformatics side, whereas sophisticated methods are available for identifying positively and negatively selected amino acid sites, no analogous tools have been developed yet for studying selection at the RNA level.

Acknowledgments

We thank Marco Vignuzzi for commentaries on the manuscript. This work was financially supported by grant

BFU2008-03978/BMC and the Ramón y Cajal program from the Spanish MICIIN.

References

- Braun MJ, Clements JE, Gonda MA. 1987. The visna virus genome: evidence for a hypervariable site in the env gene and sequence homology among lentivirus envelope proteins. *J Virol.* 61:4046–4054.
- Brower-Sinning R, Carter DM, Crevar CJ, Ghedin E, Ross TM, Benos PV. 2009. The role of RNA folding free energy in the evolution of the polymerase genes of the influenza A virus. *Genome Biol.* 10:R18.
- Carter J, Saunders V. 2007. *Virology: principles and applications.* Sussex: Wiley.
- Cawley S, Bekiranov S, Ng HH, et al. (20 co-authors). 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116:499–509.
- Contreras AM, Hiasa Y, He W, Terella A, Schmidt EV, Chung RT. 2002. Viral RNA mutations are region specific and increased by ribavirin in a full-length hepatitis C virus replication system. *J Virol.* 76:8505–8517.
- Damgaard CK, Andersen ES, Knudsen B, Gorodkin J, Kjems J. 2004. RNA interactions in the 5' region of the HIV-1 genome. *J Mol Biol.* 336:369–379.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
- Frankel AD, Young JA. 1998. HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem.* 67:1–25.
- García M, Crawford JM, Latimer JW, Rivera-Cruz E, Perdue ML. 1996. Heterogeneity in the haemagglutinin gene and emergence of the highly pathogenic phenotype among recent H5N2 avian influenza viruses from Mexico. *J Gen Virol.* 77:1493–1504.
- Hue S, Gifford RJ, Dunn D, Fernhill E, Pillay D. 2009. Demonstration of sustained drug-resistant human immunodeficiency virus type 1 lineages circulating among treatment-naive individuals. *J Virol.* 83:2645–2654.
- Ji J, Hoffmann JS, Loeb L. 1994. Mutagenicity and pausing of HIV reverse transcriptase during HIV plus-strand DNA synthesis. *Nucleic Acids Res.* 22:47–52.
- Kaempfer R. 2003. RNA sensors: novel regulators of gene expression. *EMBO Rep.* 4:1043–1047.
- Kawashima Y, Pfafferoth K, Frater J, et al. (43 co-authors). 2009. Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* 458:641–645.
- Knies JL, Dang KK, Vision TJ, Hoffman NG, Swanstrom R, Burch CL. 2008. Compensatory evolution in RNA secondary structures increases substitution rate variation among sites. *Mol. Biol. Evol.* 25:1778–1787.
- Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol.* 22:1208–1222.
- Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324:255–258.
- Le SY, Chen JH, Braun MJ, Gonda MA, Maizel JV. 1988. Stability of RNA stem-loop structure and distribution of non-random structure in the human immunodeficiency virus (HIV-1). *Nucleic Acids Res.* 16:5153–5168.
- Le SY, Chen JH, Chatterjee D, Maizel JV. 1989. Sequence divergence and open regions of RNA secondary structures in the envelope regions of the 17 human immunodeficiency virus isolates. *Nucleic Acids Res.* 17:3275–3288.
- Lindahl T, Nyberg B. 1974. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* 13:3405–3410.
- Parthasarathi S, Varela-Echavarría A, Ron Y, Preston BD, Dougherty JP. 1995. Genetic rearrangements occurring during a single cycle of murine leukemia virus vector replication: characterization and implications. *J Virol.* 69:7991–8000.
- Pedersen JS, Meyer IM, Forsberg R, Simmonds P, Hein J. 2004. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res.* 32:4925–4936.
- Pelletier J, Sonenberg N. 1988. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature* 334:320–325.
- Reuter JS, Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics.* 11:129.
- Rodríguez-Alvarado G, Roossinck MJ. 1997. Structural analysis of a necrogenic strain of cucumber mosaic cucumovirus satellite RNA in planta. *Virology* 236:155–166.
- Schinazi RF, Lloyd RM Jr, Ramanathan CS, Taylor EW. 1994. Antiviral drug resistance mutations in human immunodeficiency virus type 1 reverse transcriptase occur in specific RNA structural regions. *Antimicrob. Agents Chemother.* 38:268–274.
- Schultes E, Hraber PT, LaBean TH. 1997. Global similarities in nucleotide base composition among disparate functional classes of single-stranded RNA imply adaptive evolutionary convergence. *RNA* 3:792–806.
- Simmonds P, Smith DB. 1999. Structural constraints on RNA virus evolution. *J Virol.* 73:5787–5794.
- Simmonds P, Tuplin A, Evans DJ. 2004. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. *RNA* 10:1337–1351.
- Smit S, Knight R, Heringa J. 2009. RNA structure prediction from evolutionary patterns of nucleotide composition. *Nucleic Acids Res.* 37:1378–1386.
- Tellam J, Smith C, Rist M, Webb N, Cooper L, Vuocolo T, Connolly G, Tschärke DC, Devoy MP, Khanna R. 2008. Regulation of protein translation through mRNA structure influences MHC class I loading and T cell recognition. *Proc Natl Acad Sci USA.* 105:9319–9324.
- Tucker BJ, Breaker RR. 2005. Riboswitches as versatile gene control elements. *Curr Opin Struct Biol.* 15:342–348.
- Tuplin A, Wood J, Evans DJ, Patel AH, Simmonds P. 2002. Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. *RNA* 8:824–841.
- Voïnet O. 2009. Origin, biogenesis, and activity of plant micro-RNAs. *Cell* 136:669–687.
- Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW Jr, Swanstrom R, Burch CL, Weeks KM. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature.* 460:711–716.
- Wei X, Decker JM, Wang S, et al. (14 co-authors). 2003. Antibody neutralization and escape by HIV-1. *Nature* 422:307–312.
- Xia X, Xie Z. 2001. DAMBE: software package for data analysis in molecular biology and evolution. *J Hered.* 92:371–373.
- Xu Z, Choi J, Yen TS, Lu W, Strohecker A, Govindarajan S, Chien D, Selby MJ, Ou J. 2001. Synthesis of a novel hepatitis C virus protein by ribosomal frameshift. *EMBO J.* 20:3840–3848.
- Yoshida K, Nakamura M, Ohno T. 1997. Mutations of the HIV type 1 V3 loop under selection pressure with neutralizing monoclonal antibody NM-01. *AIDS Res Hum Retroviruses.* 13:1283–1290.
- Yu Q, Konig R, Pillai S, Chiles K, Kearney M, Palmer S, Richman D, Coffin JM, Landau NR. 2004. Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. *Nat Struct Mol Biol.* 11:435–442.