12-1-2008

Evidence of a novel RNA secondary structure in the coding region of HIV-1 pol gene.

Qi Wang
University of California at Los Angeles

Ian Barr
University of California at Los Angeles, ian.barr@dominican.edu

Feng Guo
University of California at Los Angeles

Christopher Lee
University of California at Los Angeles

https://doi.org/10.1261/rna.1252608

Evidence of a novel RNA secondary structure in the coding region of HIV-1 *pol* gene

QI WANG,[1,2] IAN BARR,[1,3] FENG GUO,[1,3] and CHRISTOPHER LEE[1,2]

[1]Molecular Biology Institute, University of California at Los Angeles, Los Angeles, California 90095, USA
[2]Center for Computational Biology, Department of Chemistry and Biochemistry, Institute of Genomics and Proteomics, University of California at Los Angeles, Los Angeles, California 90095, USA
[3]Department of Biological Chemistry, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, California 90095, USA

ABSTRACT

RNA secondary structures play several important roles in the human immunodeficiency virus (HIV) life cycle. To assess whether RNA secondary structure might affect the function of the HIV protease and reverse transcriptase genes, which are the main targets of anti-HIV drugs, we applied a series of different computational approaches to detect RNA secondary structures, including thermodynamic RNA folding predictions, synonymous variability analysis, and covariance analysis. Each method independently revealed strong evidence of a novel RNA secondary structure at the junction of the protease and reverse transcriptase genes, consisting of a 107-nucleotide region containing three stems, A, B, and C. First, RNA folding calculations by mfold and RNAfold both predicted the secondary structure with high confidence. Moreover, the same structure was predicted in a diverse set of reference sequences in HIV-1 group M, indicating that it is conserved across this group. Second, the predicted base-pairing regions displayed markedly reduced synonymous variation (approximately threefold lower than average) in a data set of 20,000 HIV-1 subtype B sequences from clinical samples. Third, independent analysis of covariation between synonymous mutations in this data set identified 10 covariant mutation pairs forming two diagonals that corresponded exactly to the sites predicted to base-pair in stems A and B. Finally, this structure was validated experimentally using selective 2′-hydroxyl acylation and primer extension (SHAPE). Discovery of this novel secondary structure suggests many directions for further functional investigation.

Keywords: HIV-1 *pol* gene; RNA secondary structure; thermodynamic prediction; covariation; synonymous variability; SHAPE

INTRODUCTION

HIV is the causative agent of AIDS, now a worldwide epidemic. One serious problem for the treatment of AIDS is HIV's ability to rapidly develop resistance to anti-retroviral drugs. The majority of FDA-approved anti-HIV drugs target the protease and the reverse transcriptase in the HIV *pol* gene (Simon et al. 2006). In order to better understand the development of drug resistance, it may be important to understand the structure and function not only of the protease and reverse transcriptase proteins, but also of the *pol* gene itself, such as possible RNA secondary structures, since these could affect its function.

A number of RNA secondary structures have been identified in different parts of the HIV genome (Paillart et al. 2002; Abbink and Berkhout 2003; Damgaard et al. 2004; Hofacker et al. 2004; Ooms et al. 2007). There are some well-studied examples, such as the *trans*-activating responsive (TAR) element at the 5′-end of the genome (Berkhout 1992), the Rev response element (RRE) in the *env* gene (Malim et al. 1989), and the *gag*/*pol* frame-shift hairpin (Parkin et al. 1992). They all have been found to play important roles in HIV transcription. In addition, it has been reported that an RNA secondary structure in the *gp120* gene facilitated recombination, creating a recombination hot spot in HIV (Moumen et al. 2001; Galetto et al. 2004). All these studies suggest that RNA secondary structure in HIV plays important roles in the viral life cycle. One study has suggested a relationship between RNA secondary structure and drug resistance mutations in HIV (Schinazi et al. 1994).

Thus, one important goal is the complete identification of all RNA secondary structures in HIV, particularly

Reprint requests to: Christopher Lee, Molecular Biology Institute, 611 Charles E. Young Drive East, 609 Boyer Hall, University of California at Los Angeles, Los Angeles, CA 90095, USA; e-mail: leec@chem.ucla.edu; fax: (310) 206-7286.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.1252608.

in regions involved in drug resistance. This requires several different kinds of analysis. Energy-based RNA folding prediction programs can give useful predictions of likely structures, but are not in and of themselves adequate evidence for a specific structure. Comparative genomic methods provide a variety of ways to test such predictions (Mathews and Turner 2006). First, comparison of many related sequences can evaluate whether regions containing predicted secondary structures are more strongly conserved than neighboring regions. Furthermore, by focusing such analysis on synonymous sites, it is possible to distinguish whether conservation is due to selection pressure on the amino acid sequence (i.e., protein function) or on the RNA sequence itself (consistent with a functionally important RNA secondary structure). Second, comparative genomics can evaluate whether the predicted secondary structure is conserved over a broader evolutionary clade. Finally, if sufficient data are available, mutation covariance analysis can directly indicate pairs of nucleotides that appear to be base-paired by identifying compensatory mutations. All of these approaches depend on having enough sequences to obtain statistically significant results.

The combination of energy-based folding and comparative genomic approaches has successfully detected RNA secondary structures in HIV. Hofacker et al. (1998) correctly identified the two well-known secondary structures TAR and RRE via a combination of thermodynamic structure prediction with phylogenetic comparison of as few as 13 full genomic sequences. The emergence of larger HIV sequence data sets provides a useful opportunity to take greater advantage of comparative genomics to identify all RNA secondary structures in HIV. Peleg et al. (2002) applied a combination of secondary structure prediction and the conservation assessment method to 178 *env* sequences and identified one novel RNA structure in the *env* gene. They also predicted another secondary structure in the *nef* gene through analyzing 106 *nef* sequences (Peleg et al. 2003).

In this study, we have applied thermodynamic prediction methods along with comparative sequence analyses of about 20,000 *pol* gene sequences, yielding strong evidence of a novel RNA secondary structure at the junction between the protease and reverse transcriptase coding regions.

RESULTS

Identification of a 107-nucleotide predicted RNA secondary structure

To search for possible RNA secondary structures in the *pol* gene, we ran RNA folding predictions for 201-nucleotide (nt) windows with a step size of 51 nt throughout the HIV-1 group M consensus gene sequence (starting from the first nucleotide of protease; see Materials and Methods), using the program RNAfold (Fig. 1; see Materials and Methods for details). To assess the significance of predictions in each window, we compared the minimal free energy (MFE) of

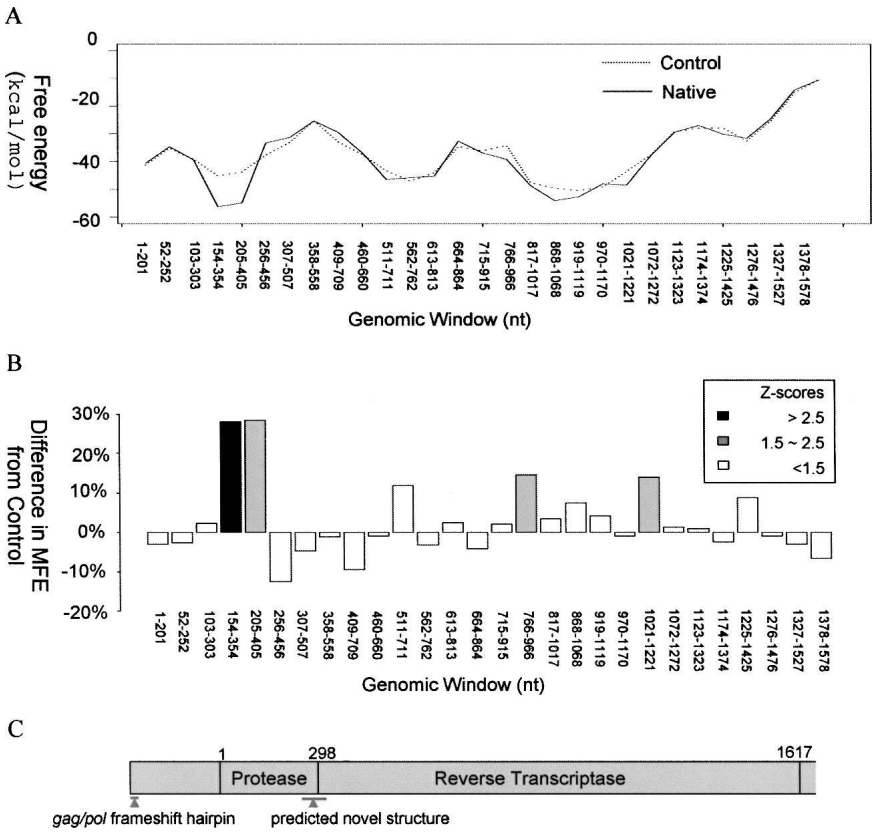

FIGURE 1. Minimal free energy (MFE) of 201-base fragments in the *pol* gene and its comparison with that of shuffled control sequences. (*A*) The *x*-axis is genomic coordinates for the 201-base fragments with respect to the first nucleotide in the protease; the *y*-axis is the MFE of the native fragment (solid line) and the mean MFE of 1000 corresponding shuffled control sequences (dotted line). (*B*) The *x*-axis is the same as in *A*; the *y*-axis is the difference in MFE of the native fragment from the mean value of 1000 shuffled control sequences. Shading indicates the significance of such difference in *Z*-score, which is the number of standard deviations from the mean of the shuffled control sequences. (*C*) Schematic diagram of the first half of the *pol* gene, up to the end of the reverse transcriptase. The genes and the positions of RNA secondary structures are marked. Coordinates are given starting from the first nucleotide in the protease, as in the rest of the article.

the consensus gene sequence versus that of shuffled control sequences. Specifically, we calculated the *Z*-score for this difference (Fig. 1B; Simmonds et al. 2004) by generating a sample of 1000 shuffled sequences using DicodonShuffle (Katz and Burge 2003; see Materials and Methods), which preserves the encoded protein sequence, codon usage, and dinucleotide composition of the original sequence.

These analyses identified a region from nucleotides 205 to 354 that yielded a predicted RNA secondary structure (Fig. 2) with statistical significance. The same stem–loop structure was predicted in two adjacent 201-nt windows (nucleotides 154–354 and 205–405) with *Z*-scores (from the sequence shuffling tests) of 2.7 and 2.3, respectively. These *Z*-scores were statistically significant, yielding *P*-values of 0.003 and 0.012, respectively. Other predictions with weaker *Z*-scores were found between nucleotides 766 and 966 ($Z = 2.1$) and 1021 and 1221 ($Z = 2.1$).

A new set of predictions using the program mfold (Zuker 1989) was generated for nucleotide windows 154–354 and 205–405. In both windows, mfold predicted an identical secondary structure of 107 nt formed by nucleotide fragments 217–323 (Fig. 2), the same structure as predicted by the RNAfold. This structure, located at the junction of the protease and the reverse transcriptase (Fig. 1C), consisted of three base-paired stems (A, B, and C) and two loops, one on the end of stem B and the other on the end of stem C.

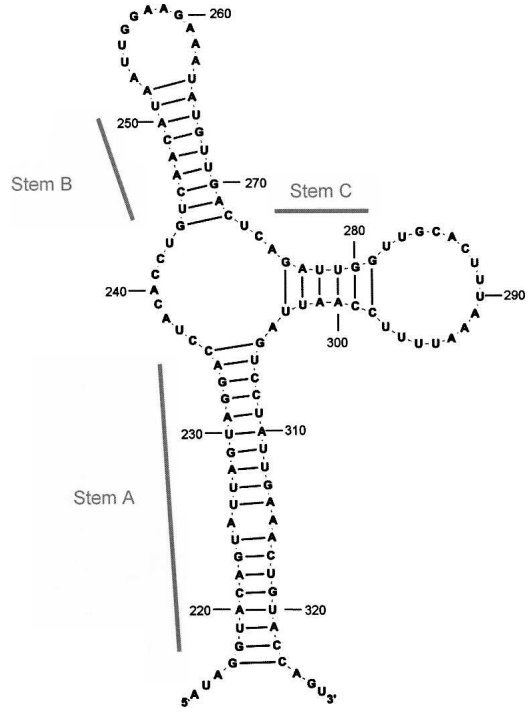

FIGURE 2. The predicted RNA secondary structure of nucleotide region 214–326 (with respect to the first nucleotide of the protease) of HIV-1 group M consensus sequence, using RNAfold with default parameters. Only base-pairings with probabilities >0.5 are shown. Stems A, B, and C are indicated.

Stem A, the longest, contained 19 base pairs (bp) of complementary sequence with only a single mismatch (at nucleotide positions 225/315). Stem B consisted of 9 bp of complementary sequence, separated by an 11-nt hairpin turn. Stem C had only 6 bp of complementary sequence.

Since the accuracy of structure prediction can be significantly improved by analysis of multiple sequences (Mathews and Turner 2006), we tested whether the predicted secondary structure is conserved over a diverse family of related HIV types. We generated secondary structure predictions for each member of the set of reference sequences representing strains of HIV group M from the Los Alamos HIV Sequence Database (see Materials and Methods). This data set consisted of a total of 37 reference sequences representing 11 distinct HIV subtypes, including subtype B. There were about four sequences for each subtype, broadly representing the subtype. For each reference, we extracted nucleotide regions 214–326 (with reference to the group M consensus in the multiple sequence alignment) and predicted the secondary structures using mfold (Zuker 1989) with default parameters. Stem A (Fig. 2) was conserved in 100% of the predicted structures; stem B was conserved in 70% of the predicted structures; stem C was conserved in 50% of the predicted structures (see Supplemental Table 1 for details). In some cases, mfold predicted multiple structures with similar energies. Overall, 100% of the reference sequences had both stems A and B in at least one of the predicted structures and 92% of the reference sequences had stem C in at least one of the predicted structures. This suggests that the predicted structure is highly conserved among different subtypes in group M. Among the three stems in the predicted structure, stem A is the most conserved, whereas stem C is the least.

We also analyzed this region using RNAz (Gruber et al. 2007), a program for predicting structurally conserved and thermodynamically stable RNA secondary structures in multiple sequence alignments. The multiple sequence alignment of the 37 group M reference sequences (see Methods and Materials) was analyzed using the program's default parameters except for a window size of 120 nt and a step size of 10 nt. RNAz predicted the same stem–loop secondary structure with the probability of 1.00. This high level of confidence suggests that the patterns of sequence conservation in HIV-1 group M are strongly consistent with this predicted RNA secondary structure.

Validation by synonymous polymorphism data from 20,000 HIV-1 sequences

The region containing the predicted RNA secondary structure is part of an open reading frame encoding the protease and reverse transcriptase proteins. Thus, conservation of nucleotide sequences can simply result from the conservation of protein sequences. Therefore, it is important

to ask whether conservation in this region might simply be due to amino acid selection pressure (i.e., selection on the protein sequence), as opposed to conservation of the proposed RNA structure. Since synonymous mutations do not alter the amino acid sequence and thus are unaffected by amino acid selection pressure, measurements of conservation using only synonymous substitutions (such as ds, Ks) can distinguish specific evidence of selection on RNA structure and function from the more typical pattern of amino acid selection pressure (Xing and Lee 2006). The obvious question is whether the predicted base-pairing regions (i.e., stems A, B, and C) show markedly lower synonymous substitution rates than surrounding regions of the gene. Very large data sets of clinical HIV-1 sequences make it possible to answer this question by measuring synonymous variability for each codon position.

We computed a measure of synonymous variability from a data set of about 20,000 HIV-1 subtype B sequences (Chen et al. 2004). Following the previous analysis of Gog et al. (2007), synonymous variability was calculated for each codon position as the normalized mean pairwise distance (MPD) (raw data shown in Supplemental Fig. 1; see Materials and Methods for details). To ensure sufficient diversity of the data and minimize potential bias from phylogenetic effects, we had excluded sequences with <2% nucleotide diversity, so that in the final set including 20,000 sequences any two sequences were >2% different from each other. Over the 450 codon positions analyzed in the *pol* gene, the average MPD score was 0.206.

In order to reveal regions with suppressed synonymous variability, we calculated the moving average of the MPD score over a sliding window of 10 codons (Gog et al. 2007). The windowing analysis revealed five regions in the *pol* gene with unusually low synonymous substitution rates (Fig. 3A, I–V). Strikingly, regions IV (MPD value of 0.09) and V (MPD value of 0.07) corresponded directly to the complementary sequences of stem A in the predicted RNA secondary structure (Fig. 2). By contrast, the low synonymous variability of region I resulted from the fact that the first 13 codons of protease overlap with the reading frame of *gag*, the previous gene, and thus synonymous sites in this region of the *pol* gene were actually nonsynonymous sites (which, typically, have much lower variability, due to amino acid selection pressure) in the *gag* gene. We have not investigated the possible causes of region II or III.

We next examined in detail the correspondence between synonymous variability of individual sites and the predicted stem–loop structure (Fig. 3B). Both halves of stem A displayed unusually low levels of synonymous variability (MPD values of 0.02–0.10). Strikingly, the one mismatch in the predicted stem A structure, formed by positions 225 and 315, had dramatically higher levels of synonymous variability (MPD 0.32 and 0.18, respectively, similar to the average level of MPD in *pol*) than the predicted base-paired

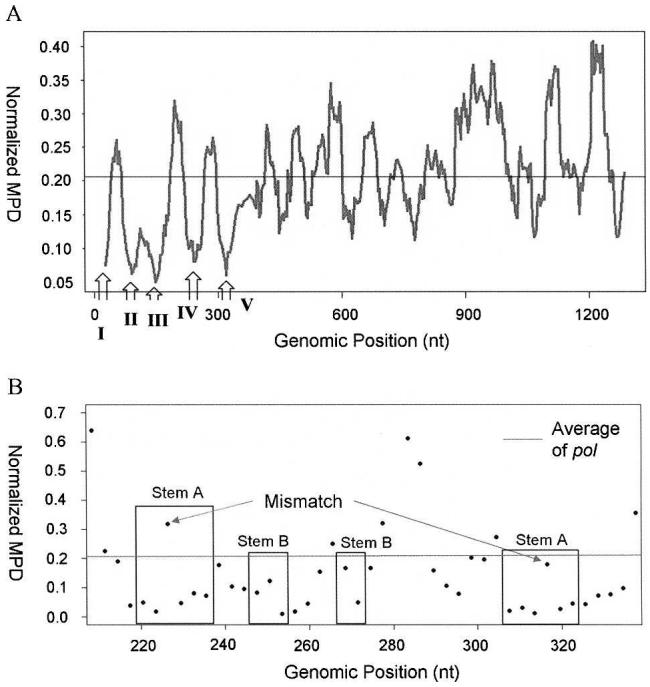

FIGURE 3. Synonymous variability in the coding region of the *pol* gene. (*A*) The moving average of synonymous variability in windows of 10 codons with a step size of one codon. The variability of each codon is measured by normalized mean pairwise distance (MPD). The *x*-axis is average genomic coordinates of codons in the scanning window with respect to the first nucleotide in the protease; the *y*-axis is the averaged MPD score in the window. The five low synonymous variability regions, regions I–V, are indicated by arrows. The horizontal line indicates the average MPD score of the *pol* region. (*B*) The synonymous variability of the codon positions in the sequence region that folds into the predicted structure. The *x*-axis is the genomic coordinates as in *A*; the *y*-axis is the MPD score. The horizontal line indicates the average MPD score of the *pol* region. The base-pairing regions of stems A and B are indicated by boxes.

positions in stem A (mean MPD 0.04 with standard deviation 0.02). We evaluated the statistical significance of the low synonymous variability of the positions predicted to base-pair in stem A using the Wilcoxon rank test. This result was strongly significant (*P*-value of 3×10^{-4}). Stem B also appeared to have reduced synonymous variability, although this result was weaker (*P*-value of 0.059), consistent with the short length of stem B, and stem C showed no significant reduction (*P*-value of 0.90). We also analyzed the synonymous variability of the two weaker predictions of base-paired stems between nucleotides 766–966 and 1021–1221; neither was statistically significant.

HIV-1 has been shown to have a tendency to become A-rich (van Hemert and Berkhout 1995). To assess the potential bias resulting from this nucleotide sequence pressure, we have calculated the percentage of A in the 107-nt region of the *pol* gene that folds into the secondary structure and compared it with that of the entire *pol* gene. The percentage of A is 32% (34/107) for this region and 39% (1190/3015) for the full *pol* gene, showing no

significant difference (Fisher's exact test *P*-value = 0.3). Thus the data do not support A-pressure as an explanation for the low mutation rate observed in the secondary structure region.

Validation of specific base pairs by analysis of covariant substitutions

With extremely large sequence data sets, it becomes possible to search for validation of base pairing by finding pairs of mutations that show strong covariance, indicating that they are "compensatory mutations" (Chen et al. 1999). Again, in a protein-coding region such as *pol* it is extremely important to screen out patterns of covariance that might be due to amino acid selection pressure (i.e., selection for compensatory mutations due to their effects on protein function) (Wang and Lee 2007). Since synonymous mutations do not alter the amino acid sequence, covariance between synonymous mutations cannot be due to amino acid selection pressure. We therefore analyzed covariance between all possible pairs of synonymous mutations in *pol*, as an independent test of possible base pairings, using the same data set of 20,000 HIV-1 subtype B sequences (Chen et al. 2004).

This analysis highlights one striking feature of the predicted RNA secondary structure (Fig. 4A). By random chance, there is only a 1/3 probability that synonymous sites in one strand of an RNA stem will be base-paired to synonymous sites in the other strand. From this point of view, it is striking that synonymous sites base-paired only with synonymous sites in all three stems in the predicted RNA secondary structure (*P*-value = 0.037).

Covariance analysis yielded a sparse scatter of pairs of sites with statistically significant covariation (Fig. 4B; complete data in Supplemental Fig. 2; see Materials and Methods for details). No pattern was evident except for two clusters of strongly covarying pairs, each forming a line perpendicular to the diagonal, indicative of the antiparallel orientation required for a base-pairing hairpin structure. These two clusters corresponded directly to positions

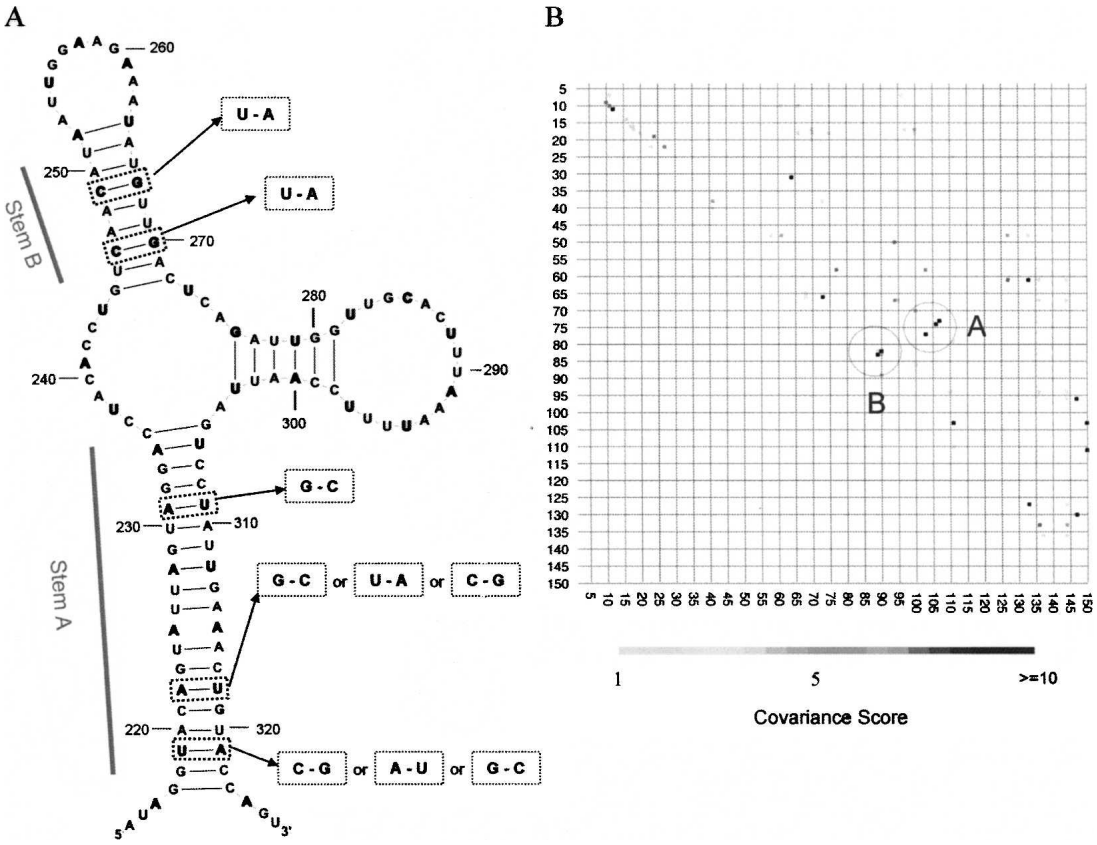

FIGURE 4. Covariant synonymous mutations in the predicted RNA secondary structure. (*A*) The predicted structure of nucleotide regions 214–326 (with respect to the first nucleotide of protease) of HIV-1 group M consensus sequence, using RNAfold with default parameters. Only base-pairings with probabilities >0.5 are shown. The covariant substitutions are given in the boxes for each pair of covariant sites. The third nucleotide in each codon is in bold. Stems A and B are indicated. (*B*) The covariation map shows the strength of covariance of synonymous substitutions for each codon position pair. The *x*- and *y*-axes are coordinates for the codon positions, starting from the first codon in the protease. For a given codon position pair, the highest covariance value for any pair of synonymous mutations at the two positions is displayed on the map. The two clusters of strongly covarying sites are highlighted in circles and labeled with the corresponding stem index A or B.

predicted to be base-paired in stems A and B, respectively. Not only did the covarying pairs match precisely the positions predicted to pair, but 9 out of 10 pairs of nucleotide mutations formed standard Watson–Crick base pairs (see Fig. 4A and Supplemental Table 2). In addition, two pairs of sites in stem A each contained three pairs of covariant substitutions (Table 1). These covariance results were highly statistically significant, by both the chi-square test and Fisher's exact test (P-value $< 10^{-10}$; see Materials and Methods). The existence of multiple covariant sites in stems A and B provides compelling independent evidence for the precise base pairs predicted by the RNA folding calculations.

To assess the significance of the synonymous covariation results as evidence for the predicted secondary structure, we performed several statistical tests. Considering all possible pairs of synonymous mutations within the 1.4-kb region analyzed in this study, there are a total of 124,663 possible synonymous mutation pairs that would satisfy Watson–Crick base pairing. Thus, to obtain a significance level of 0.01 after the Bonferroni correction, we set a P-value cutoff of 10^{-7} for any specific synonymous mutation pair. Only 36 pairs (0.03% of the total) met this criterion for significant covariation. By contrast, of the only 16 possible pairs of complementary synonymous mutations at the pairing sites in the three stems of the proposed secondary structure (Fig. 4A), nine (or 56% of the total) were actually detected by this criterion as showing significant covariation. Indeed, the P-values for these pairs were strong (P-value ranging from 10^{-8} to 10^{-42}). These data provide very strong validation for the predicted secondary structure, because they do not merely confirm the general region of the predicted stem, but specifically validate the precise base-pairing of the predicted secondary structure. This is a statistically strong result. The likelihood of obtaining these results by random chance is only 5.4×10^{-29} (see hypergeometric test, Materials and Methods).

Covariant substitutions were observed not only in HIV-1 subtype B sequences but also in subtype H. In the group M consensus sequence, there was one paired C and G at nucleotide positions 249 and 267, respectively, which was located in stem B. This pair of sites remained complementary after substitutions at both positions in subtype H, to nucleotides U and A, respectively. The covariant substitutions provide independent confirmation of the secondary structure inferred from both the synonymous variability analysis and the free energy-based folding.

Experimental validation using the SHAPE assay

We probed the secondary structure using selective 2′-hydroxyl acylation and primer extension (SHAPE). SHAPE exploits the difference in reactivity between paired and unpaired nucleotides toward *N*-methylisatoic anhydride (NMIA) (Merino et al. 2005; Wilkinson et al. 2006). NMIA modifies an RNA at a 2′ hydroxyl, and the modification blocks subsequent primer extension using a reverse transcriptase. The products of reverse transcription are examined using sequencing gels, enabling direct readout of the modification pattern. Increased gel band intensities in the NMIA-treated lane (+) versus the untreated lane (−) identify RNA residues that are reactive and thus likely unpaired.

A 113-nt sequence fragment of a HIV-1 subtype A1 reference (see Materials and Methods) that contains the proposed structure was synthesized using in vitro transcription and was subjected to SHAPE analysis (Fig. 5). We observed high reactivity (>50%) for the nucleotides in the loops of the B and C stems, the junctions between stems A and B and stems C and A, as well as the 5′ and 3′ flanking regions. In contrast, stems A and B show little NMIA reactivity, indicating that they are strongly base-paired as in the predicted secondary structure. Stem C, the shortest

TABLE 1. Evidence of covariation for pairs of nucleotide sites

		Nucleotide 219			
	Observed	A	G	C	U (wt)
	G	0	0	**50**	166
Nucleotide 321	U	**4**	1	0	3
	C	0	**4**	0	3
	A (wt)	42	76	93	20,584
		Nucleotide 222			
	Observed	G	U	C	A (wt)
	A	0	**7**	0	50
Nucleotide 318	G	0	0	**7**	56
	C	**12**	0	0	26
	U (wt)	54	6	13	20,798

The numbers indicating the standard Watson–Crick base pairs are shown in bold.
(wt) the consensus.

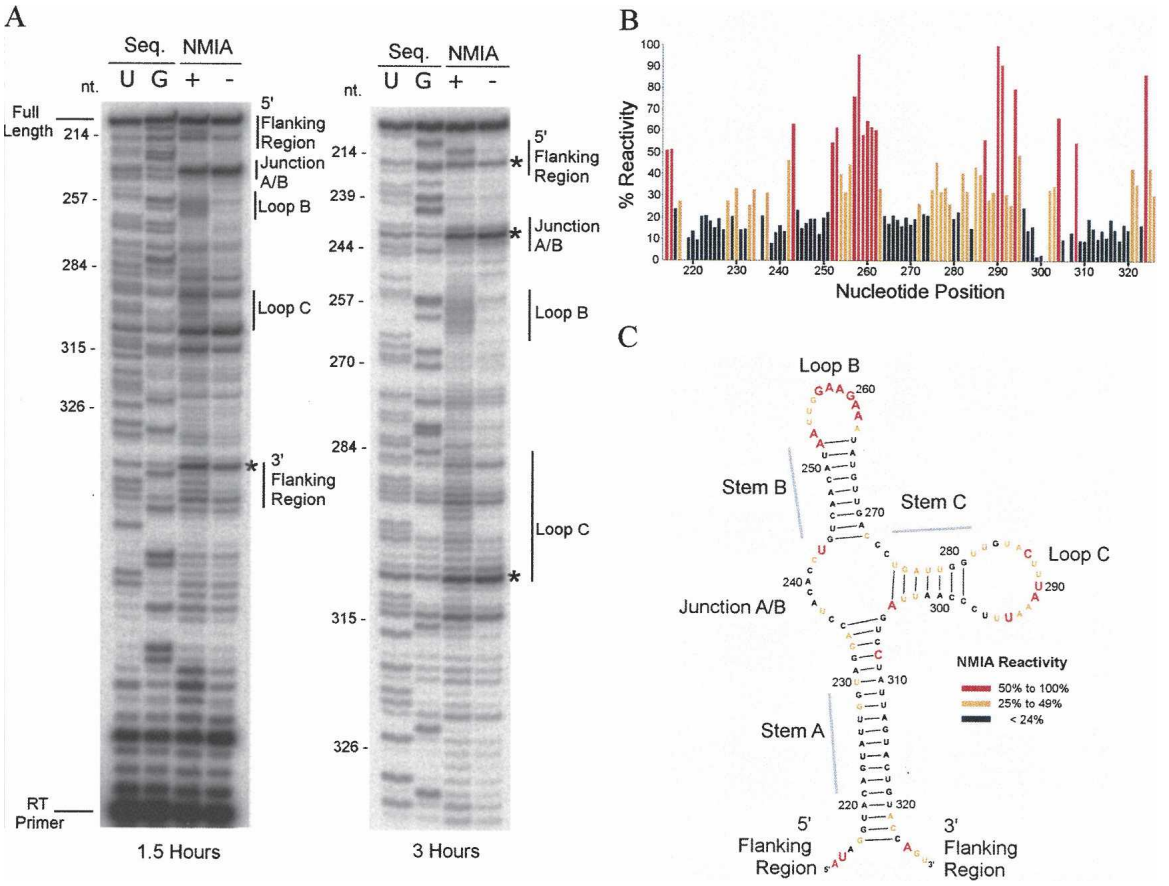

FIGURE 5. SHAPE analysis of the proposed HIV secondary structure. (*A*) Sequencing polyacrylamide gels showing SHAPE analysis of the RNA after running 1.5 h (*left*) and 3 h (*right*) at 60 W, as was necessary to resolve bands on the 5′ and 3′ ends. The lanes marked U and G are sequencing lanes, in which ddATP and ddCTP were added to the reverse transcription to cause chain termination at U and G sites. NMIA lanes (+) and (−) correspond to incubation with 60 mM NMIA for 45 min at 37°C and a control with NMIA omitted. The flexible segments are visible as regions of increased NMIA modification in the (+) lane as compared to the (−) lane. The asterisks indicate the pauses of reverse transcription even in the absence of NMIA, likely caused by the helices A and C. (*B*) SHAPE modification intensities, determined from the gels in *A*. SHAPE intensities of residues 214–309 were determined from the 3-h gel, and 310–326 from the 1.5 h gel. Bars show the amounts of modification at each position relative to the most highly modified nucleotide. (*C*) SHAPE modification intensities mapped onto the secondary structure, showing regions of heavy modification at unpaired nucleotides. The structure was generated using RNAfold. Only base pairings with probabilities >0.5 are shown. Residues with intensities between 25% and 49% and 50% and 100% are shown in orange and red, respectively. The sequence used in this analysis was retrieved from a reference genome of HIV-1 subtype A1 (GenBank accession number: U51190) from the Los Alamos HIV Sequence Database.

helix, shows modest to low degrees of modification. In addition, the reverse transcriptase tends to pause at the 3′ end of strong secondary structures. Under the conditions we used, strong pauses in reverse transcription were observed for stems A and C even in the absence of NMIA (Fig. 5A, asterisks), consistent with our predicted secondary structure. Therefore, the SHAPE data support the novel 3-stem RNA structure that we have proposed.

DISCUSSION

Detection of an RNA secondary structure using evolutionary data

The first indication of this novel secondary structure came from our previous studies of HIV mutation covariance in clinical HIV samples (Chen et al. 2004; Wang and Lee 2007), which identified the unusual pattern of synonymous mutation covariance shown in Figure 4. The clear ''diagonal'' pattern in the covariance matrix, the remarkable conservation of Watson–Crick base-pairing by compensatory mutations, and the obvious sequence complementarity of these regions immediately indicated an RNA secondary structure.

Subsequent secondary structure prediction using both thermodynamic prediction methods as well as methods that exploit the large amount of comparative sequence data—around 20,000 clinical HIV samples—yielded a consistent folding prediction across the HIV-1 group M family. The thermodynamic prediction programs mfold and RNAfold both predicted a secondary structure matching the base pairing seen in our covariance data. In

addition, the folding free energy of the native sequence was significantly lower than those of shuffled control sequences (Fig. 1), indicating a statistically significant result. Moreover, stems A and B persisted in the predictions of all the sequences throughout the diverse set of reference sequences of HIV-1 group M. Also, the base-paired regions of the predicted structure were highly conserved at the nucleotide level, even at synonymous sites, which have no effect on the protein sequence (Fig. 3). In the coding regions of viral genomes, reduced synonymous mutation rates have been found to be a signature for nucleotide secondary structures (Simmonds and Smith 1999) or other nucleotide regulatory sequences (Gog et al. 2007). It should be noted that our bioinformatics analysis focuses on positive-strand RNA, because the HIV life cycle includes only positive-strand RNA, and the biological relevance of the negative strand is not clear.

It should be emphasized that these evolutionary data provide evidence not only of a secondary structure, but also of its functional importance to the reproductive fitness of HIV. We first predicted this secondary structure based on measurements of selection pressure against synonymous mutations and synonymous mutation covariance. We observed that specific synonymous mutations occur far less frequently than would be expected under a neutral model, indicating that individual viruses containing these mutations reproduced much less successfully than viruses with comparable mutations at other sites. If our selection pressure measurements had a purely spurious basis, predictions based upon them should also be spurious. The fact that their predicted base-paired regions have been validated by the SHAPE experiment suggests that they represent real selection pressures. All of the available evidence indicates that even a single mutation disrupting this base-pairing will have much lower reproductive success (as measured by our data from wild HIV populations) than comparable synonymous mutations at other nearby sites. This is evidence of an important function, but with no suggestion of what that function might be.

There are a variety of computational methods (Hofacker et al. 2002; Knudsen and Hein 2003; Pedersen et al. 2004) that combine covariant substitution models with the thermodynamic prediction methods. These methods are superior if we have only a small set of sequences, where the evidence from covariant substitutions can hardly infer the secondary structure by itself. In our analysis, the large sequence data set enables the covariance information to serve as independent evidence for the predicted structure, predicting individual base pairs with significant *P*-values.

Experimental detection of RNA secondary structure using SHAPE

SHAPE appears to be an effective and efficient new approach for detecting RNA secondary structure. Based on the principle that unpaired regions are more accessible to NMIA acylation (Merino et al. 2005), it exploits the fact that acylated nucleotides will terminate reverse transcription to yield a result that can be easily read out like a sequencing gel. Originally developed by Weeks and coworkers, it has been applied to a variety of problems including studies of tRNA and the dimerization domain of a retroviral genome (Badorrek and Weeks 2005; Wilkinson et al. 2005), yielding results consistent with established methods of structure analysis. Another study has demonstrated the accuracy of a SHAPE analysis of the telomerase RNA using NMR (Chen et al. 2006). Recently, SHAPE was applied to a different section of the HIV-1 genome, the first 906 nt of the 5′ noncoding region (Wilkinson et al. 2006), revealing several secondary structural elements. These studies indicate that SHAPE is a robust and widely applicable technique.

There is a striking level of agreement between the SHAPE results, which measure nucleotide accessibility and conformation versus the synonymous mutation and covariance results, which measure evolutionary selection pressure. They both give strong evidence supporting stems A and B (Figs. 3–5). In contrast, at stem C, the most flexible stem structure according to SHAPE, there is no evidence of conservation in comparative sequence analyses. First, stem C shows no significant reduction in synonymous variability vs. neighboring regions (*P*-value = 0.9). Second, no significant mutation covariation was detected in stem C (Fig. 4). Third, among the 37 reference sequences of HIV-1 group M, the stem C is the least conserved stem in the predicted structure (Supplemental Table 1).

Role of the RNA secondary structure

What is the possible biological function of this secondary structure? This is largely unknown. One possibility in the literature is that it might assist recombination. The predicted structure is similar to a secondary structure whose role in recombination in HIV has been well studied (Galetto et al. 2004). They both contain an ~20-bp-long stem at the base (Fig. 2, stem A versus S1 by Galetto et al. [2004]); they both contain a 7-9 bp stem (Fig. 2, stem B versus S2 by Galetto et al. [2004]) less than 10 nt downstream of stem A or S1, with a bubble on the end of the stem B or S2. In addition, the well-studied structure contains an 18-nt bubble (Galetto et al. 2004, L1) downstream of stem S2, while downstream of the corresponding stem B there is also a bubble of 16 nt at the end of stem C. In contrast to the similarity between our predicted structure and the structure known to assist recombination, the predicted structure is less similar to the secondary structures in HIV with other functions, such as TAR (Berkhout 1992) and the *gag/pol* frame-shift hairpin (Parkin et al. 1992), both of which are single stem–loop structures, as well as RRE (Malim et al. 1989), which is a

more complicated structure with compounded stem–loop structures.

Based on the example of the S1/S2 structure (Galetto et al. 2004), one might predict that a recombination hot spot could be found within our proposed structure (Galetto et al. 2006). Recombination breakpoints have been found in nucleotide region 230–330, i.e., in the sequence that folds to the predicted structure, between the *pol* gene sequences from different subtypes, such as between subtypes A and B, A and D, and B and D, as well as between B and F (Quarleri et al. 2004; Yang et al. 2004; Sa Filho et al. 2005). However, the existence of breakpoints in this region is insufficient to define a recombination hot spot. Laboratory experiments need to be designed specifically to test the hypothesis that the recombination frequency in this region is higher than that of the neighboring regions (Moumen et al. 2001).

However, even if this explanation were correct, it seems unlikely that recombination alone could explain the very strong levels of selection pressure observed in this sequence region, implying the possibility of other, more important functions. An obvious first step would be to measure the effect of these synonymous mutations on HIV reproductive success in standard laboratory assays. If this secondary structure is shown to be important for the viral life cycle in the laboratory, it will be both interesting and possible to dissect its functional importance experimentally.

MATERIALS AND METHODS

HIV-1 subtype B sequence data

This data set contained just subtype B sequences, mostly from patients under antiretroviral drug treatment (Chen et al. 2004; Pan et al. 2007). These sequences cover 450 codons, including the whole protease (99 codons) and the first 351 codons of the reverse transcriptase. Multiple sequence alignments and mutation detection were performed as previously described (Chen et al. 2004). To ensure sufficient diversity of the data and minimize potential bias from phylogenetic effects, we excluded sequences with less than 2% nucleotide diversity. After filtering, there are 20,042 sequences available, which have been included in this analysis.

HIV-1 consensus sequences and reference sequences

The consensus sequence of HIV-1 group M was downloaded from the Los Alamos HIV Sequence Database version August 2004. Aligned HIV-1 group M reference sequences were downloaded from HIV-1 Subtype Reference Alignments in the version 2007 of Los Alamos HIV Sequence Database.

Thermodynamic prediction, free energy calculation, and base-pair probabilities

RNAfold in the Vienna package version 1.6.1 and mfold version 3.2 (http://frontend.bioinfo.rpi.edu/) were used to predict structures using the default parameters. RNAfold was also used to measure the MFE for each sequence with its default parameters. It predicts the free energy of the most stable RNA structure for a given sequence. The base-pair probabilities were calculated by RNAfold as well (McCaskill 1990).

Sequence shuffling tests

All sequence randomization was carried out using the Dicodon-Shuffle algorithm (Katz and Burge 2003), which retains the dinucleotide composition at the (3,1), (1,2), and (2,3) positions as well as the encoded amino acid sequence and codon usage of the native coding sequence. The program was downloaded from the authors' website at ftp://hollywood.mit.edu.

Measure of synonymous variability

At each amino acid position, we calculated the normalized mean pairwise distance (MPD) as described (Gog et al. 2007). The MPD, as a measure of sequence variability, is the sum of all individual pairwise distances divided by the number of pairs [n sequences give $n(n-1)/2$ possible sequence pairs]. At each amino acid position, we only analyzed codons coding the most frequent amino acid, so that the difference between any two codons is synonymous. Hence, these MPD scores do not reflect the amino acid-level selection (Simmonds and Smith 1999). To take codon bias into consideration, the MPD score is normalized by the expected MPD given the distribution of codons for that amino acid in the whole segment (Gog et al. 2007). At positions with invariant tryptophan and methionine codons, the expected MPDs are zero. Positions with these amino acids are not included in our moving average calculation.

Measure of covariance

We used Fisher's exact test (Fisher 1922; Agresti 1992) to test for nonrandom associations between mutation α at position X and mutation β at position Y, by computing the P-value for the two-sided test using the 2×2 contingency table: $N_{X\alpha Y\beta}$, $N_{X\alpha Y0}$, $N_{X0Y\beta}$, and N_{X0Y0}. $N_{X\alpha Y\beta}$ is the number of samples that have mutation α at position X and also mutation β at position Y; $N_{X\alpha Y0}$ is the number of samples that have mutation α at position X but no mutation at position Y; $N_{X0Y\beta}$ is the number of samples that have no mutation at position X and have mutation β at position Y; N_{X0Y0} is the number of samples that have mutations at neither position. We computed the odds ratio, its confidence interval (95% two-sided), and the P-value using the fisher.test function from the statistical software package R. The lower-bound estimate for the strength of covariation, based on a 95% confidence interval, is defined as the covariance score.

Covariation map

The covariation map was derived as described (Wang and Lee 2007). Only statistically significant mutation pairs ($P < 10^{-8}$ for a single pair by Fisher's exact test, yielding a significance level of 0.003 for each pair after the Bonferroni correction for all possible synonymous mutation pairs) were included in our analysis. For a given codon position pair, the strongest covariance score for any pair of synonymous mutations at the two positions was displayed in the map.

Chi-square test and Fisher's exact test for covariant sites involving multiple pairs of mutations

The chi-square test (the chisq.test function from the statistical software package R) was used to test for nonrandom associations

between synonymous mutations at position *X* with those at position *Y*. The contingency table used in the chi-square test has two variables for positions *X* and *Y*, respectively, each including all possible synonymous mutations at that position. The number in the contingency table where *X* is α and *Y* is β indicates the number of samples having synonymous difference α at position *X* and synonymous difference β at position *Y*.

Fisher's exact test (the fisher.test function from the statistical software package R) was used on the same contingency table, except the columns and rows associated with the wild type (consensus), because the program cannot handle the large numbers in these columns and rows.

Hypergeometric significance test for enrichment of covariant mutation pairs

Assuming there are *M* pairs of strongly covariant mutations out of a total of *N* pairs and we observed *m* pairs of strongly covariant mutations in a subset containing *n* pairs, the statistical significance that the subset is enriched with strongly covarying pairs is inferred by the probability that *m* or more strongly covarying pairs are observed in the subset under a random model. Under the random model, if we randomly draw *n* pairs from the total set of *N* pairs, the probability that *m* of them are strongly covarying pairs follows the Hypergeometic distribution (dhyper function in the statistical software package R).

SHAPE analysis

Cloning and transcription

The SHAPE method (Merino et al. 2005; Wilkinson et al. 2006) was used to probe unpaired regions in the predicted secondary structure. A 113-nt sequence fragment, nucleotides 1833–1945 of a reference genome of HIV-1 subtype A1 (GenBank accession number: U51190), was synthesized using in vitro transcription. The RNA was cloned into the pUC19 vector between EcoRI and XbaI restriction sites. The construct contains, in addition to the HIV sequence, a 5′ linker and a 3′ linker containing the RT primer binding site (Wilkinson et al. 2006). This cloned plasmid was linearized with XbaI and used in run-off transcription by T7 RNA Polymerase (Milligan and Uhlenbeck 1989). Transcriptions were precipitated in ethanol and purified on a 6% denaturing polyacrylamide gel (29:1 acrylamide:bis-acrylamide). The RNA was resuspended in 0.5× TE buffer (5 mM Tris at pH 7.5, 0.5 mM EDTA).

Radiolabeling

Reverse transcription primer from IDT (Integrated DNA Technologies) was labeled using γ^{32}P ATP and polynucleotide kinase. The labeled, full-length primer was purified on a 20% denaturing polyacrylamide gel. The primer was ethanol precipitated before use in reverse transcription.

NMIA modification

We followed the method of Wilkinson et al. (2006) with a few modifications. We annealed 12 pmol of RNA at 95°C for 3 min, then chilled it at 4°C for 1 min. We then added 3.33× annealing buffer (333 mM HEPES at pH 8.0, 20 mM $MgCl_2$, 333 mM NaCl) and incubated at 37°C for 20 min. The annealed RNA was treated with 6 mM NMIA in DMSO or DMSO alone for 45 min at 37°C, then ethanol precipitated and reverse transcribed.

Sequencing gel running and analysis

Sequencing reactions were run on a 12% denaturing polyacrylamide gel. The sequencing gels were run at 60 W for the indicated times and dried at 80°C for 60 min, and then exposed to a storage phosphor screen for 72 h. The gel images were produced using a Typhoon phosphorimager (GE Healthcare). Analysis and quantification were carried out using the SAFA program (Das et al. 2005). Band intensities in the NMIA (−) lane were subtracted from the (+) lane in order to determine the contribution of NMIA to the signal in the (+) lane, and the highest degree of NMIA modification was set to 100%.

SUPPLEMENTAL DATA

Supplemental material can be found at http://www.rnajournal.org.

ACKNOWLEDGMENTS

We thank Alex Alekseyenko, Mijeong Kang, Aron Yoffe, and the three anonymous reviewers for helpful comments and discussions on the manuscript; Art Poon for helping to calculate the covariation controlling for phylogeny; and Frederick Bibollet-Ruche, Matteo Negroni, Frank Maldarelli, and Paul Sharp for helpful discussions on the potential function of the structure. This work was supported by grants from the NIH (U54 RR021813) and DOE (DE-FC02-02ER63421), a Dreyfus Foundation Teacher-Scholar Award to C. L., and UCLA dissertation year fellowship and UCLA AIDS Institute fellowship (UCLA AIDS Institute, and the UCLA Center for AIDS Research [AI28697]) to Q.W.

Received January 24, 2008; accepted September 24, 2008.

REFERENCES

Abbink, T.E. and Berkhout, B. 2003. A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon. *J. Biol. Chem.* **278:** 11601–11611.

Agresti, A. 1992. A survey of exact inference for contingency tables. *Stat. Sci.* **7:** 131–177.

Badorrek, C.S. and Weeks, K.M. 2005. RNA flexibility in the dimerization domain of a γ retrovirus. *Nat. Chem. Biol.* **1:** 104–111.

Berkhout, B. 1992. Structural features in TAR RNA of human and simian immunodeficiency viruses: A phylogenetic analysis. *Nucleic Acids Res.* **20:** 27–31.

Chen, Y., Carlini, D.B., Baines, J.F., Parsch, J., Braverman, J.M., Tanda, S., and Stephan, W. 1999. RNA secondary structure and compensatory evolution. *Genes Genet. Syst.* **74:** 271–286.

Chen, L., Perlina, A., and Lee, C.J. 2004. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J. Virol.* **78:** 3722–3732.

Chen, Y., Fender, J., Legassie, J.D., Jarstfer, M.B., Bryan, T.M., and Varani, G. 2006. Structure of stem–loop IV of *Tetrahymena* telomerase RNA. *EMBO J.* **25:** 3156–3166.

Damgaard, C.K., Andersen, E.S., Knudsen, B., Gorodkin, J., and Kjems, J. 2004. RNA interactions in the 5′ region of the HIV-1 genome. *J. Mol. Biol.* **336:** 369–379.

Das, R., Laederach, A., Pearlman, S.M., Herschlag, D., and Altman, R.B. 2005. SAFA: Semiautomated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. *RNA* **11:** 344–354.

Fisher, R.A. 1922. On the interpretation of $\chi(2)$ from contingency tables, and the calculation of *P. J. R. Stat. Soc. [Ser A]* **85:** 87–94.

Galetto, R., Moumen, A., Giacomoni, V., Veron, M., Charneau, P., and Negroni, M. 2004. The structure of HIV-1 genomic RNA in the gp120 gene determines a recombination hot spot in vivo. *J. Biol. Chem.* **279:** 36625–36632.

Galetto, R., Giacomoni, V., Veron, M., and Negroni, M. 2006. Dissection of a circumscribed recombination hot spot in HIV-1 after a single infectious cycle. *J. Biol. Chem.* **281:** 2711–2720.

Gog, J.R., Afonso Edos, S., Dalton, R.M., Leclercq, I., Tiley, L., Elton, D., von Kirchbach, J.C., Naffakh, N., Escriou, N., and Digard, P. 2007. Codon conservation in the influenza A virus genome defines RNA packaging signals. *Nucleic Acids Res.* **35:** 1897–1907.

Gruber, A.R., Neubock, R., Hofacker, I.L., and Washietl, S. 2007. The RNAz web server: Prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Res.* **35:** W335–W338.

Hofacker, I.L., Fekete, M., Flamm, C., Huynen, M.A., Rauscher, S., Stolorz, P.E., and Stadler, P.F. 1998. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res.* **26:** 3825–3836.

Hofacker, I.L., Fekete, M., and Stadler, P.F. 2002. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319:** 1059–1066.

Hofacker, I.L., Stadler, P.F., and Stocsits, R.R. 2004. Conserved RNA secondary structures in viral genomes: A survey. *Bioinformatics* **20:** 1495–1499.

Katz, L. and Burge, C.B. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.* **13:** 2042–2051.

Knudsen, B. and Hein, J. 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.* **31:** 3423–3428.

Malim, M.H., Hauber, J., Le, S.Y., Maizel, J.V., and Cullen, B.R. 1989. The HIV-1 rev *trans*-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature* **338:** 254–257.

Mathews, D.H. and Turner, D.H. 2006. Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.* **16:** 270–278.

McCaskill, J.S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29:** 1105–1119.

Merino, E.J., Wilkinson, K.A., Coughlan, J.L., and Weeks, K.M. 2005. RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127:** 4223–4231.

Milligan, J.F. and Uhlenbeck, O.C. 1989. Synthesis of small RNAs using T7 RNA polymerase. *Methods Enzymol.* **180:** 51–62.

Moumen, A., Polomack, L., Roques, B., Buc, H., and Negroni, M. 2001. The HIV-1 repeated sequence R as a robust hot-spot for copy-choice recombination. *Nucleic Acids Res.* **29:** 3814–3821.

Ooms, M., Abbink, T.E., Pham, C., and Berkhout, B. 2007. Circularization of the HIV-1 RNA genome. *Nucleic Acids Res.* **35:** 5253–5261.

Paillart, J.C., Skripkin, E., Ehresmann, B., Ehresmann, C., and Marquet, R. 2002. In vitro evidence for a long range pseudoknot in the 5′-untranslated and matrix coding regions of HIV-1 genomic RNA. *J. Biol. Chem.* **277:** 5995–6004.

Pan, C., Kim, J., Chen, L., Wang, Q., and Lee, C. 2007. The HIV positive selection mutation database. *Nucleic Acids Res.* **35:** D371–D375.

Parkin, N.T., Chamorro, M., and Varmus, H.E. 1992. Human immunodeficiency virus type 1 *gag-pol* frameshifting is dependent on downstream mRNA secondary structure: Demonstration by expression in vivo. *J. Virol.* **66:** 5147–5151.

Pedersen, J.S., Meyer, I.M., Forsberg, R., Simmonds, P., and Hein, J. 2004. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res.* **32:** 4925–4936.

Peleg, O., Brunak, S., Trifonov, E.N., Nevo, E., and Bolshoy, A. 2002. RNA secondary structure and squence conservation in C1 region of human immunodeficiency virus type 1 env gene. *AIDS Res. Hum. Retroviruses* **18:** 867–878.

Peleg, O., Trifonov, E.N., and Bolshoy, A. 2003. Hidden messages in the nef gene of human immunodeficiency virus type 1 suggest a novel RNA secondary structure. *Nucleic Acids Res.* **31:** 4192–4200.

Quarleri, J.F., Rubio, A., Carobene, M., Turk, G., Vignoles, M., Harrigan, R.P., Montaner, J.S., Salomon, H., and Gomez-Carrillo, M. 2004. HIV type 1 BF recombinant strains exhibit different pol gene mosaic patterns: Descriptive analysis from 284 patients under treatment failure. *AIDS Res. Hum. Retroviruses* **20:** 1100–1107.

Sa Filho, D.J., Sanabani, S., Diaz, R.S., Munerato, P., Brunstein, A., Fusuma, E., Sabino, E.C., and Janini, L.M. 2005. Analysis of full-length human immunodeficiency virus type 1 genome reveals a variable spectrum of subtypes B and f recombinants in Sao Paulo, Brazil. *AIDS Res. Hum. Retroviruses* **21:** 145–151.

Schinazi, R.F., Lloyd Jr., R.M., Ramanathan, C.S., and Taylor, E.W. 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.

Simmonds, P. and Smith, D.B. 1999. Structural constraints on RNA virus evolution. *J. Virol.* **73:** 5787–5794.

Simmonds, P., Tuplin, A., and Evans, D.J. 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.

Simon, V., Ho, D.D., and Abdool Karim, Q. 2006. HIV/AIDS epidemiology, pathogenesis, prevention, and treatment. *Lancet* **368:** 489–504.

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. *J. Mol. Evol.* **41:** 132–140.

Wang, Q. and Lee, C. 2007. Distinguishing functional amino acid covariation from background linkage disequilibrium in HIV protease and reverse transcriptase. *PLoS One* **2:** e814. doi: 10.1371/journal.pone.0000814.

Wilkinson, K.A., Merino, E.J., and Weeks, K.M. 2005. RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(Asp) transcripts. *J. Am. Chem. Soc.* **127:** 4659–4667.

Wilkinson, K.A., Merino, E.J., and Weeks, K.M. 2006. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): Quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protocols* **1:** 1610–1616.

Xing, Y. and Lee, C. 2006. Can RNA selection pressure distort the measurement of Ka/Ks? *Gene* **370:** 1–5.

Yang, C., Li, M., Shi, Y.P., Winter, J., van Eijk, A.M., Ayisi, J., Hu, D.J., Steketee, R., Nahlen, B.L., and Lal, R.B. 2004. Genetic diversity and high proportion of intersubtype recombinants among HIV type 1-infected pregnant women in Kisumu, western Kenya. *AIDS Res. Hum. Retroviruses* **20:** 565–574.

Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* **244:** 48–52.