ARTICLES

Architecture and secondary structure of an entire HIV-1 RNA genome

Joseph M. Watts[1], Kristen K. Dang[2], Robert J. Gorelick[5], Christopher W. Leonard[1], Julian W. Bess Jr[5], Ronald Swanstrom[3], Christina L. Burch[4] & Kevin M. Weeks[1]

Single-stranded RNA viruses encompass broad classes of infectious agents and cause the common cold, cancer, AIDS and other serious health threats. Viral replication is regulated at many levels, including the use of conserved genomic RNA structures. Most potential regulatory elements in viral RNA genomes are uncharacterized. Here we report the structure of an entire HIV-1 genome at single nucleotide resolution using SHAPE, a high-throughput RNA analysis technology. The genome encodes protein structure at two levels. In addition to the correspondence between RNA and protein primary sequences, a correlation exists between high levels of RNA structure and sequences that encode inter-domain loops in HIV proteins. This correlation suggests that RNA structure modulates ribosome elongation to promote native protein folding. Some simple genome elements previously shown to be important, including the ribosomal *gag-pol* frameshift stem-loop, are components of larger RNA motifs. We also identify organizational principles for unstructured RNA regions, including splice site acceptors and hypervariable regions. These results emphasize that the HIV-1 genome and, potentially, many coding RNAs are punctuated by previously unrecognized regulatory motifs and that extensive RNA structure constitutes an important component of the genetic code.

Genomes of all single-stranded RNA viruses contain internal structures fundamental to viral replication and host defence evasion. Important viral RNA structures include internal ribosome entry sites, packaging signals, pseudoknots, transfer RNA mimics, ribosomal frameshift motifs, and *cis*-regulatory elements[1,2]. In the human immunodeficiency virus (HIV), RNA structures activate transcription, initiate reverse transcription, facilitate genomic dimerization, direct HIV packaging, manipulate reading frames, regulate RNA nuclear export, signal polyadenylation, and interact with viral and host proteins[2–6]. These RNA regulatory motifs have been identified by focusing on the 5′ and 3′ untranslated regions plus a few internal sequences. Most potential regulatory structures within viral RNA genomes, including in ~85% of the HIV-1 genome, are uncharacterized. This raises the possibility that new categories of RNA structure-mediated regulation remain to be identified.

The HIV-1 genome is primarily a coding RNA and contains nine open reading frames that produce 15 proteins[2,3]. The Gag polyprotein precursor is proteolytically processed to generate the matrix (MA), capsid (CA), nucleocapsid (NC) and p6 proteins. The Gag-Pol polyprotein contains protease (PR), reverse transcriptase (RT) and integrase (IN). The *env* gene encodes a 30-amino-acid signal peptide (SP), gp120 and gp41. Other sequences encode auxiliary proteins (Fig. 1a, grey boxes). Inside virions, HIV genomic RNA is found as a non-covalent dimer, is 5′ capped and 3′ polyadenylated, and is annealed to a host $tRNA^{Lys3}$ molecule[2]. Viral proteins, especially nucleocapsid, chaperone the folding of HIV RNA[7].

Whole-genome structure analysis

To develop an accurate view of RNA structure in the full-length genome, we analysed authentic genomic RNA extracted from HIV-1 virions. Our gentle purification maintained both previously characterized secondary structures and the few known RNA tertiary structures. For example, the host $tRNA^{Lys3}$ was bound to the genome[2] and a pseudoknot in the 5′ untranslated region (UTR)[6,8] remained stably formed. The RNA was sufficiently intact to act as a template for primer extension reactions spanning the entire genome (Supplementary Table 1 and Methods).

High-throughput selective 2′-hydroxyl acylation analysed by primer extension (SHAPE)[6,9–11] was used to chemically interrogate local nucleotide flexibility at 99.4% of the 9,173 nucleotides in the NL4-3 HIV-1 RNA genome. 1-methyl-7-nitroisatoic anhydride (1M7) preferentially acylates conformationally flexible nucleotides at the ribose 2′-OH position[9,10]. The resulting 2′-*O*-adducts are detected as stops to primer extension using fluorescently labelled primers and capillary electrophoresis[6,10] (Fig. 3a) and are quantified by whole-trace Gaussian integration[11] (Fig. 3b). SHAPE measurements are reproducible between independent biological replicates ($R^2 = 0.75$; Supplementary Fig. 1). SHAPE reactivities are highly sensitive to local nucleotide flexibility and disorder, but are insensitive to solvent accessibility[9,12] (Supplementary Fig. 2).

SHAPE reactivities therefore provide direct model-free information about the overall level of structure, or architecture, for any RNA. The median SHAPE reactivity varies markedly across the HIV-1 genome (Fig. 1b, dark blue line). Regions with median reactivities below 0.25 indicate domains with substantial base-paired secondary RNA structure, whereas median SHAPE reactivities of 0.5 and greater indicate regions of largely unstructured nucleotides.

We also assessed HIV-1 genome structure by examining evolutionary information contained in nucleotide and amino acid variation to assign a pairing probability at each nucleotide[13,14]. This algorithm does not use chemical reactivity or thermodynamic information, and thus infers RNA structure using information that is orthogonal to SHAPE.

We identify at least 10 'structured' regions that exhibit both low SHAPE reactivity and high pairing probability (Fig. 1b, compare dark

[1]Department of Chemistry, [2]Department of Biomedical Engineering, [3]Linenberger Cancer Center, [4]Department of Biology, University of North Carolina, Chapel Hill, North Carolina 27599-3290, USA. [5]AIDS and Cancer Virus Program, SAIC-Frederick, Inc., NCI-Frederick, Frederick, Maryland 21702-1201, USA.

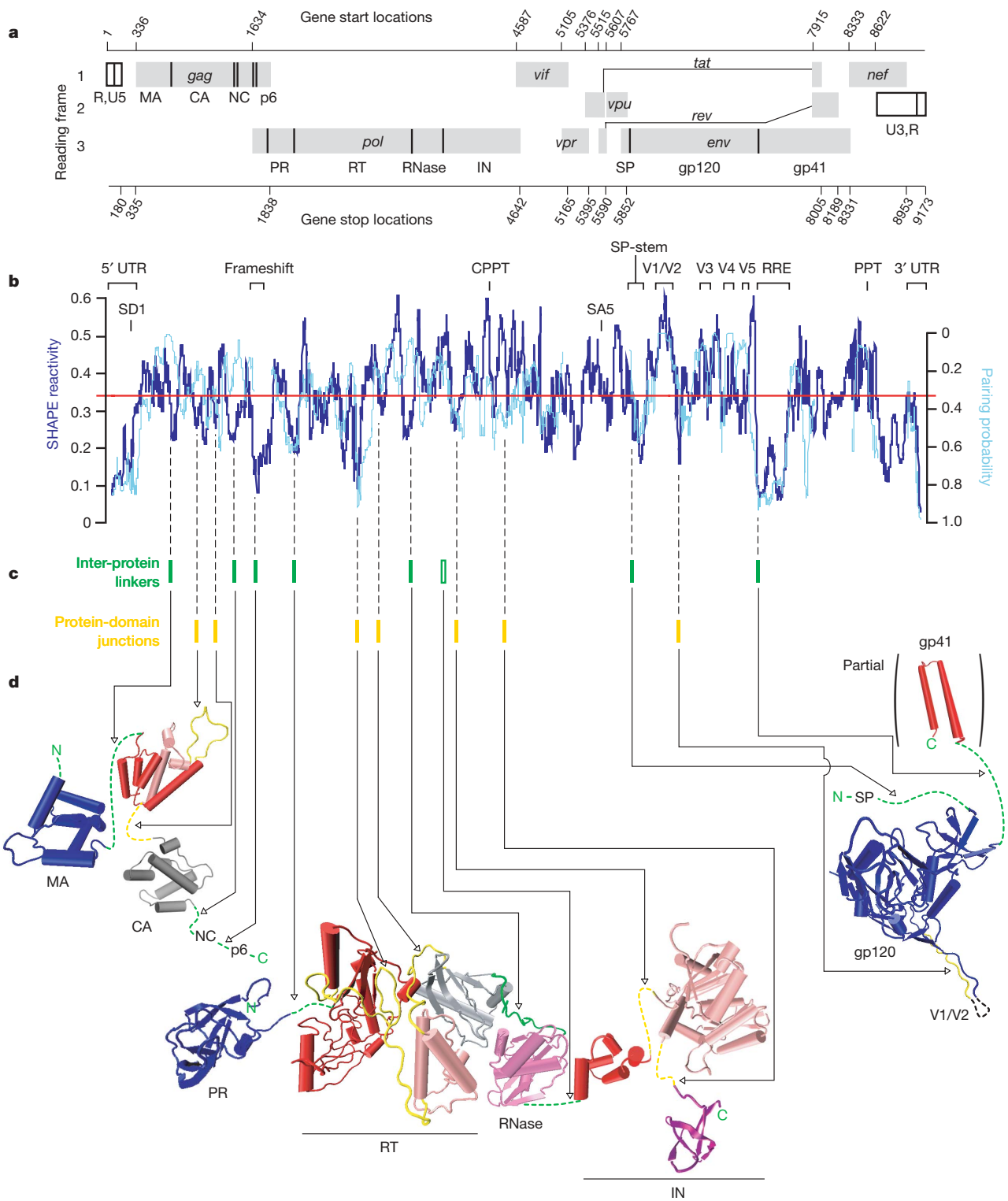

Figure 1 | Organization, extent of RNA structure, and relationship to protein structure for an HIV-1 genome. **a**, HIV-1 genome organization. Protein coding regions are shown as grey boxes; polyprotein-domain junctions are depicted as solid vertical lines. Gene start and end sites are numbered according to NL4-3. CA, capsid; IN, integrase; MA, matrix; NC, nucleocapsid, PR, protease; RT, reverse transcriptase; SP, signal peptide. **b**, Comparison of median SHAPE reactivities (dark blue line) and evolutionary pairing probabilities (cyan line). Medians are calculated using a 75-nucleotide window. The global median (0.34) is depicted as a red line. Pairing probability is not reported for regions encoding overlapping reading frames. PPT, polypurine tract; CPPT, central PPT. **c**, Inter-protein linkers in polyprotein precursors and the unstructured peptide loops that link protein domains are indicated with green and yellow bars, respectively. The single inter-protein linker that is not encoded by a region of highly structured RNA (at the RNase H–integrase junction) is shown with an open green bar. **d**, Comparison of domain structures for the large HIV proteins with the structure of the encoding RNA. Polyprotein linkers are green; inter-domain loops are yellow; folded protein domains are blue, red, light magenta, purple and grey (Supplementary Table 2).

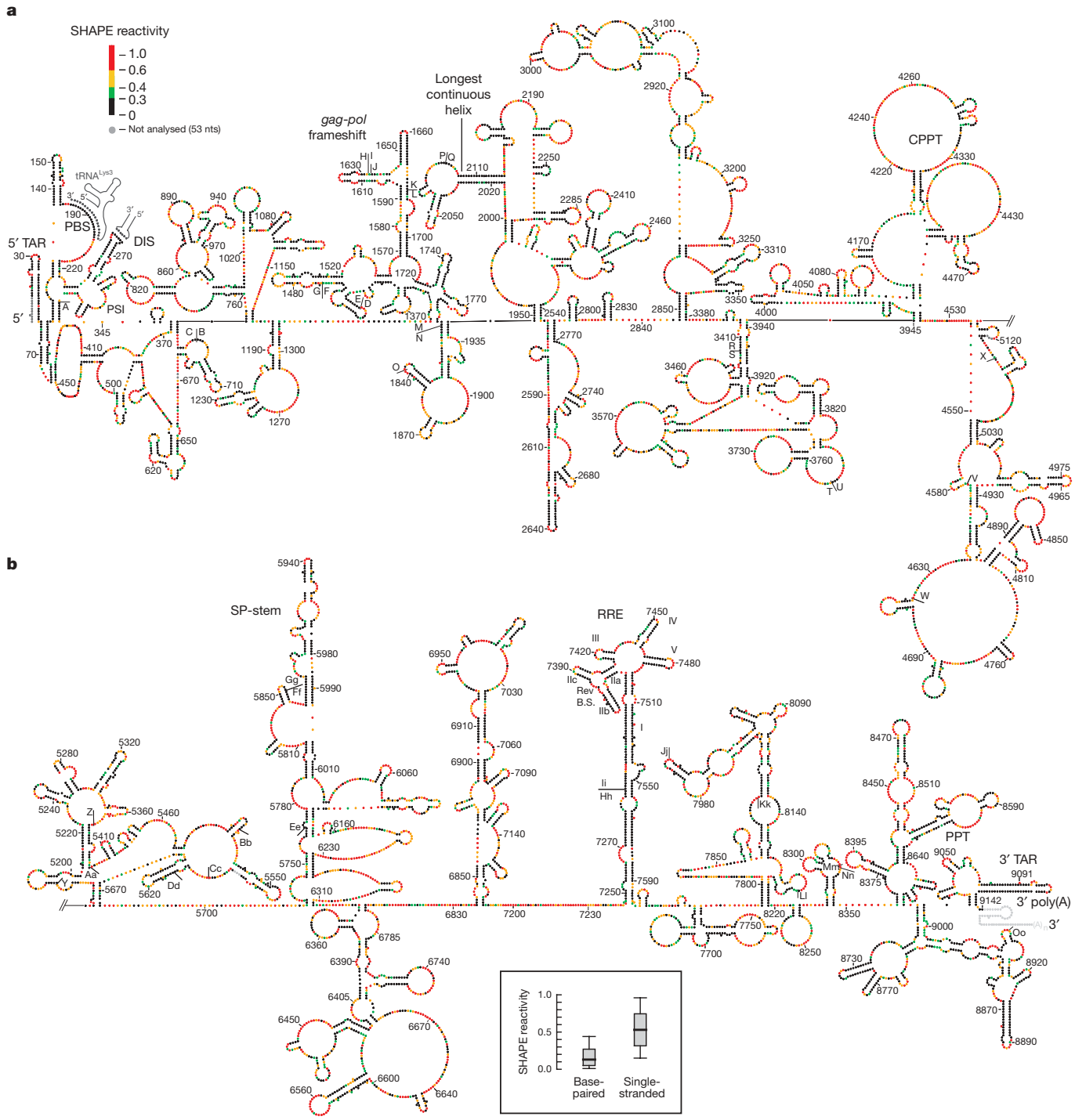

Figure 2 | Structure of the HIV-1 NL4-3 genome. The 5′ (**a**) and 3′ (**b**) genome halves are shown. Nucleotides are coloured by their absolute SHAPE reactivities (see scale in **a**). Every nucleotide is shown explicitly as a sphere; base pairing is indicated by adjacent parallel orientation of the spheres. Protein domains are identified by letters; TF, transframe peptide; nts, nucleotides. Important structural landmarks are labelled explicitly. Full nucleotide identities and pairings are provided in the Supplementary Information (Supplementary Fig. 7). Intermolecular base pairs involving the $tRNA^{Lys3}$ primer and the genomic dimer are shown in grey. Inset shows a box plot illustrating SHAPE reactivities for single-stranded versus paired nucleotides in the model. Median reactivities are indicated by bold horizontal lines; the large box spans the central 50% of the reactivity data.

blue and cyan traces). This group includes the 5′ UTR and the Rev responsive element (RRE), which are known HIV regulatory elements (Fig. 1b). However, most of these highly structured and evolutionarily conserved elements have not been characterized previously. These regions include the protease–reverse transcriptase junction, domains in the reverse transcriptase, integrase, and Vif open reading frames, an element 3′ of the Env signal peptide, and the *nef* 3′-UTR region.

We also identify at least seven ‘unstructured’ regions, extending over 200–600 nucleotides, with high SHAPE reactivities and low

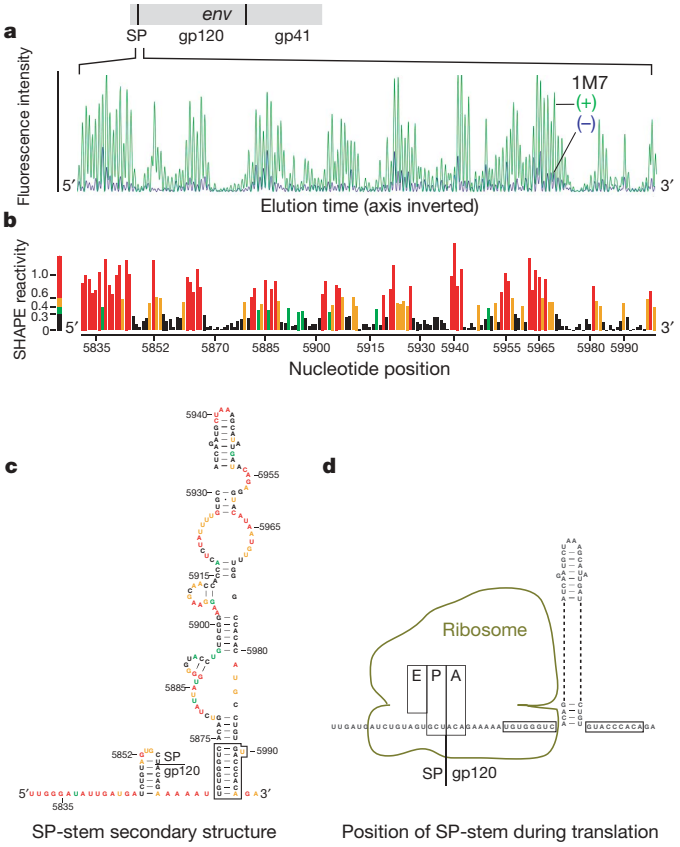

Figure 3 | SHAPE analysis of the signal peptide–gp120 region. a, Processed capillary electrophoresis trace showing intensity versus position for the (+) and (−) reagent reactions. **b**, Histogram of integrated and normalized SHAPE reactivities as a function of nucleotide position. The SHAPE reactivity scale shown here is used consistently throughout this work. **c**, RNA secondary structure model for the signal peptide pause site stem. **d**, Location of the signal-peptide stem relative to the eukaryotic ribosome at the pause site. Base pairs disrupted when the ribosome is at the pause site are boxed.

pairing probabilities. These include the RNase H coding domain, variable domains (V*x*) in gp120, and the polypurine tract (Fig. 1b). On a smaller scale, the consensus sequences for the highly used splice site acceptors are also unstructured (Supplementary Fig. 3). There are four regions of apparent disagreement in the level of RNA structure, having high pairing probabilities and high SHAPE reactivities (one each in the reverse transcriptase, RNase, integrase and gp41 coding regions). This small group may reflect sequence conservation that is not accounted for by the evolutionary model[13], or may form critical structures at an alternative stage of the viral replication cycle.

RNA structure encodes protein structure

We first evaluated whether global RNA genome structure is linked to protein structure. HIV-1 produces three major classes of messenger RNA. The 9 kilobase (kb) class encodes Gag and Gag-Pol and is identical to the packaged genomic RNA analysed here except, as an mRNA, it is not dimerized at its 5′ end[2]. There are very few differences in the SHAPE reactivity of dimeric and monomeric RNAs at the 5′ end of the genome[6]. Thus, genome structures outside of the dimerization region will correlate closely to the mRNA that encodes Gag and Gag-Pol. The most abundant 4 kb *env* mRNA is generated by splicing nucleotide 288 (SD1, the major splice donor) to nucleotide 5522 (termed the SA5 site)[15]. SA5 is followed by an unstructured genome region (Fig. 1a, b). Thus, RNA structures identified in the *env* coding region probably exist in the spliced mRNA that encodes Env. Structures for the 1.8-kb class of mRNAs, which generate Tat and Rev, cannot be predicted using the genomic RNA because discontinuous segments are joined in the final mRNA.

The Gag, Gag-Pol and Env polyprotein precursors are synthesized roughly as beads on a string, and the constituent proteins are liberated by proteolytic cleavage[2,3] (Fig. 1a, d). Eight inter-protein peptides link the HIV proteins (Fig. 1c, green bars). The RNA sequences that encode these spacer peptide linkers in Gag (at the matrix–capsid, capsid–nucleocapsid and nucleocapsid–p6 junctions), Pol (protease–reverse transcriptase and reverse transcriptase–RNase H junctions) and Env (signal peptide–gp120 and gp120–gp41 junctions) all (except the RNase–integrase junction) have SHAPE reactivities that are much lower than the median (Fig. 1b). RNA sequences that encode these inter-protein peptide linkers are more highly structured than 95.2% of randomly selected regions in the genome (Supplementary Fig. 4a).

Domains in the individual HIV-1 proteins—capsid, reverse transcriptase and integrase—are also linked by unstructured peptide elements, and each domain junction is encoded by an RNA region of low SHAPE reactivity (compare yellow bars in Fig. 1c with dark blue trace in Fig. 1b). Protein loops encoded by RNA regions with low SHAPE reactivity include the cyclophilin loop and the linker between the amino- and carboxy-terminal domains in capsid, both loops that link independently folded domains in reverse transcriptase, and the eight and nine amino acid loops linking the three domains in integrase (Fig. 1d, in yellow). These protein-domain junctions are more highly structured than 88.9% of randomly selected equivalent-length regions in the genome (Supplementary Fig. 4b).

In contrast to the other large HIV proteins, domains in gp120 (termed inner, outer and bridging sheet) are not structurally autonomous. The C-terminal 35 residues of gp120 weave from the outer to the inner domain, and the bridging sheet is comprised of residues that are 315 positions distant[16]. Junctions between domains in gp120 are also not encoded by highly structured RNA, suggesting that gp120 folding is not linked to RNA structure in the same way as for other HIV proteins because its constituent domains are not structurally independent.

The recurring pattern of structure, conspicuously located near or after autonomously folding protein coding domains, is consistent with a model in which HIV protein structure is encoded in its RNA at two distinct levels. The first is the linear relationship between RNA and protein primary sequences. In the second level, higher-order RNA structure directly encodes protein tertiary structure, because unstructured protein loops are derived from highly structured RNA elements. Many proteins appear to fold during translation[17], highly structured RNA slows and causes ribosomal pausing during translation[18,19], and changes in the extent of local RNA structure modulate protein activity[20]. Together, these observations suggest that attenuation of ribosome elongation by highly structured RNA at protein-domain junctions facilitates native folding of HIV proteins by allowing time for domains to fold independently during translation.

This model makes the clear prediction that ribosome pause sites should occur preferentially in the highly structured regions of an HIV-1 RNA that encode protein junctions. We tested this idea using a toeprinting experiment, in which ribosome processivity is inhibited by cycloheximide and sites preferentially occupied by the ribosome are detected as stops to primer extension in an *in vitro* translation reaction[21]. Ribosome pause sites are statistically overrepresented at the matrix–capsid and capsid–nucleocapsid junctions in Gag and at the sequences encoding the cyclophilin loop in capsid (Supplementary Fig. 5). Conversely, ribosome pause sites are underrepresented in flanking, but unstructured, regions of the HIV RNA ($P = 0.018$). These experiments thus strongly support the model that mRNA structure over a region spanning 60–100 nucleotides specifically modulates ribosome processivity at protein-domain junctions.

RNA secondary structure model for HIV-1

Comprehensive SHAPE reactivity information can also be used to determine a nucleotide-resolution secondary structure model for the entire NL4-3 HIV-1 genome (Fig. 2). SHAPE reactivities are converted

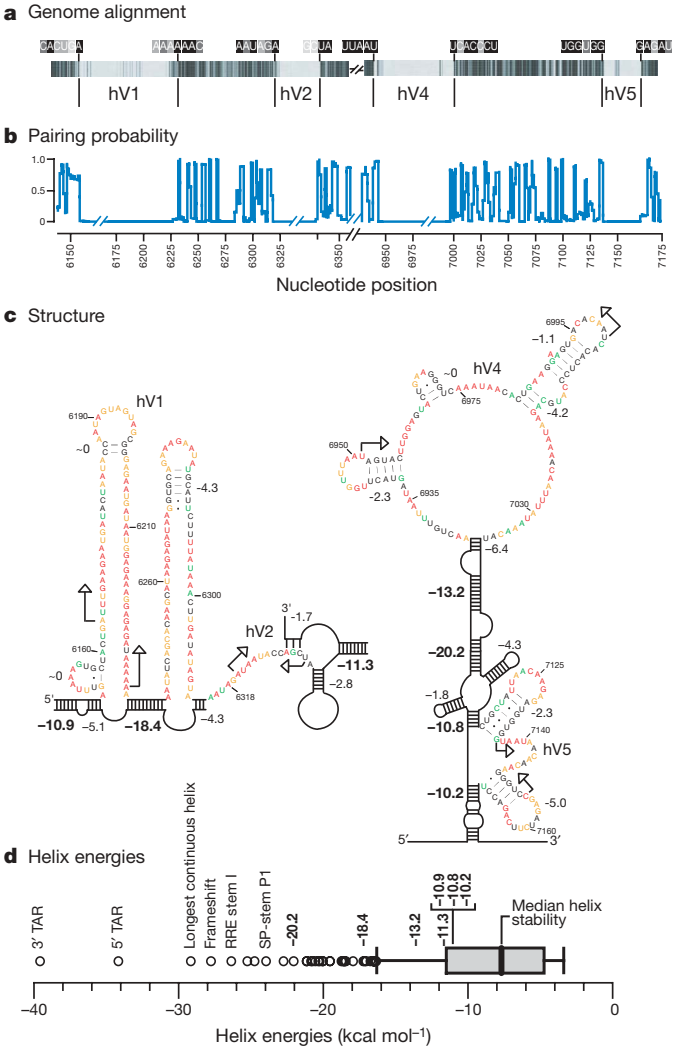

Figure 4 | RNA structure in Env hypervariable regions. **a**, Schematic sequence alignment for group M reference sequences[14] at the Env hypervariable regions (hV1, hV2, hV4 and hV5). Nucleotides are represented as vertical bars; light grey and black indicate low versus universal conservation, respectively. **b**, Evolutionary pairing probabilities. Breaks indicate extensive nucleotide insertions and deletions among the group M consensus sequences. **c**, RNA structures at the hypervariable coding regions hV1, hV2, hV4 and hV5. Calculated free energies are shown for each helix (in kcal mol^{-1}); energies for anchoring helices proposed to function as structural insulators are emphasized in bold. **d**, Distribution of helix stabilities in the HIV genome shown in a box plot representation. Whiskers illustrate 1.5-times the interquartile range, and circles emphasize helices of exceptionally high stability. Free-energy changes for proposed insulating helices are in bold; other significant helices are labelled.

to free-energy change terms and used to constrain a thermodynamic folding algorithm[22,23]. The final result is a thermodynamically favoured structural model highly reflective of the experimental SHAPE data, at single nucleotide resolution. For example, most nucleotides assigned to single-stranded regions are reactive towards SHAPE (Fig. 2, red, orange and green nucleotides), whereas base-paired nucleotides are predominantly unreactive (Fig. 2, black nucleotides and inset). For a full discussion of SHAPE-directed RNA folding and the fundamental correctness of this model, see the Methods.

The HIV-1 genome is less structured than ribosomal RNA but, similarly, contains independent RNA folding domains that extend from the overall genomic backbone. These domains include both small stem-loops plus roughly 21 large and complexly folded structures (Fig. 2). Although many genome regions are highly structured, only seven helices span a complete turn of an 11-base pair (bp) RNA duplex. The largest paired region, devoid of bulges, is the structured RNA element that bridges the coding junction between the reverse transcriptase and RNase H folding domains (Fig. 1). This helix is 19-bp long, contains a non-canonical G-A base pair (Fig. 2a, nucleotides 2015–2033/2103–2121), and is thus shorter than the 30-bp length competent to induce the interferon response[24].

The HIV-1 genome structural model provides a robust starting point for identifying previously unrecognized functional elements and long-range RNA interactions. SHAPE reactivities describe a well-formed stem 3′ to the signal-peptide coding region in the Env protein (Fig. 3c). This stem (the signal-peptide stem) is evolutionarily conserved (Fig. 1b), reinforcing an important biological role. The signal recognition particle (SRP) binds the nascent Env signal peptide and translocates the cytoplasmic ribosome elongation complex to the rough endoplasmic reticulum, where translation of gp120 and gp41 continue[25].

RNA-induced translational pausing occurs as the ribosome unwinds highly structured RNA, typically located 6–7 nucleotides downstream of the A-site[18]. The signal-peptide stem will be exactly in this conformation when the final tRNA^{Ala} from the signal peptide and the first tRNA^{Thr} of gp120 are in the P- and A-sites (Fig. 3d, boxed nucleotides). We infer that ribosomal attenuation or pausing at the signal-peptide stem provides more time for SRP recruitment and subsequent translocation of the elongation complex to the endoplasmic reticulum.

The SHAPE-constrained secondary structure is also informative for previously identified regulatory motifs. In HIV-1, *pro* and *pol* gene products are translated when the ribosome undergoes a −1 register shift from the *gag* to the *pol* reading frames. Frameshifting occurs at a slippery sequence (UUUUUUA) and is enhanced by a downstream RNA structure. These elements are typically drawn as a single-stranded slippery sequence and a 12-bp stem-loop[26]. Direct analysis of intact genomic RNA shows that the *gag-pol* frameshift signal is one component (identified here as P3) of a three-helix structure (Fig. 2 and Supplementary Fig. 6a). The slippery sequence pairs to form one of the three helices (P2). These two helices are stabilized by an anchoring helix (P1) that creates this discrete structural element (Supplementary Fig. 6a). This three-helix junction structure is conserved among HIV-1 group M sequences (Supplementary Fig. 6b).

Most RNA viruses require a complex pseudoknotted structure to induce ribosomal frameshifting[27]. The three-helix junction may function, in part, to slow translation before the ribosome encounters P3, facilitating the prerequisite pause necessary for frameshifting. The three-helix junction model may also explain why changing the slippery site to sequences that allow alternative tRNA pairing and enhance frameshifting in other RNA viruses eliminates frameshifting in HIV-1 (ref. 28). In the SHAPE-directed model, changes to the slippery sequence compromise base pairing in the conserved P2 helix (Supplementary Fig. 6).

Unstructured motifs and insulator helices

Analysis of the HIV-1 genome structure supports a role for RNA structures in sequestering unstructured regions. Five variable domains (V1–V5; see Fig. 1a, b) in the Env surface protein, gp120, account for much of the genetic diversity in HIV-1 (ref. 14) and are a critical component of the viral host evasion strategy. Four of these domains are hypervariable (hV1, hV2, hV4 and hV5) and exhibit large amino acid insertions and deletions between viral isolates[14].

Sequences encoding hypervariable domains are internally unstructured and are bordered by evolutionarily conserved and stable RNA structures (Fig. 4a, b). For example, hypervariable region hV1 is encoded by RNA sequences with high SHAPE reactivities and is flanked by two stable helices (with free energies of −10.9 and −18.4 kcal mol^{-1}, Fig. 4c). Similar patterns are evident in the other hypervariable regions (Fig. 4c). Some hypervariable regions, especially hV4, contain internal helices with non-trivial free energies; however, these helices are not evolutionarily conserved (Fig. 4b)

and are much less stable than the flanking helices that have stabilities in the 10–20 kcal mol^{-1} range (Fig. 4c). These helices are also highly stable relative to the distribution of duplex stabilities over the entire genome (Fig. 4d).

Collectively, these data suggest that RNA sequences encoding length polymorphisms in *env* are segregated from the rest of the genome by stable helices that function as structural insulators. The observed organization of hypervariable regions is thus well suited, first, to accommodate extensive substitutions, insertions or deletions, and second, to prevent these regions from forming non-functional base-pairing interactions with adjacent regulatory motifs, which include the 3′ splice site acceptors and the RRE.

Perspective

Structural analysis of a complete HIV-1 genome reveals that this RNA is punctuated by previously unrecognized, but readily identifiable and evolutionarily conserved, RNA structures. Most genome regions with low SHAPE reactivities are associated with a regulatory function (Fig. 1). SHAPE may be generally useful for identifying new regulatory elements in large RNAs. All of these elements represent hypotheses and starting points that we hope will stimulate further detailed examination. Our discovery that the peptide loops that link independently folded protein domains are encoded by highly structured RNA indicates that these and probably other mRNAs encode protein structure at a second level beyond specifying the amino acid sequence. In this view, higher-order RNA structure directly encodes protein structure, especially at domain junctions. The extraordinary density of information encoded in the structure of large RNA molecules (Figs 1, 2 and 4d) represents another level of the genetic code, one which we understand the least at present. This work makes clear that there is much to be discovered by broad structural analyses of RNA genomes and intact mRNAs.

METHODS SUMMARY

Full length HIV-1 genomic RNA was gently purified from NL4-3 virions (GenBank accession AF324493). The RNA was equilibrated in a native buffer (50 mM HEPES (pH 8.0), 200 mM potassium acetate (pH 8.0), 3 mM $MgCl_2$) at 37 °C for 15 min and treated with 1M7 (ref. 10). Sites of 2′-hydroxyl modification were identified over read lengths spanning several hundred nucleotides using 31 primer extension reactions resolved by fluorescence-detected capillary electrophoresis[6,11]. Pairing probabilities were determined using RNA-Decoder[13] and secondary structure models were developed by incorporating SHAPE reactivities as a pseudo-free-energy change term, in conjunction with nearest-neighbour parameters, in an accurate thermodynamics-based prediction algorithm[22,23].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 May; accepted 22 June 2009.

1. Cann, A. J. *Principles of Molecular Virology* Ch. 2–5 (Elsevier, 2005).
2. Coffin, J. M., Hughes, S. H. & Varmus, H. E. *Retroviruses* (Cold Spring Harbor Laboratory Press, 1997).
3. Frankel, A. D. & Young, J. A. HIV-1: fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67,** 1–25 (1998).
4. Damgaard, C. K., Andersen, E. S., Knudsen, B., Gorodkin, J. & Kjems, J. RNA interactions in the 5′ region of the HIV-1 genome. *J. Mol. Biol.* **336,** 369–379 (2004).
5. Goff, S. P. Host factors exploited by retroviruses. *Nature Rev. Microbiol.* **5,** 253–263 (2007).
6. Wilkinson, K. A. *et al.* High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol.* **6,** e96 (2008).
7. Levin, J. G., Guo, J., Rouzina, I. & Musier-Forsyth, K. Nucleic acid chaperone activity of HIV-1 nucleocapsid protein: critical role in reverse transcription and molecular mechanism. *Prog. Nucleic Acid Res. Mol. Biol.* **80,** 217–286 (2005).
8. Paillart, J. C., Skripkin, E., Ehresmann, B., Ehresmann, C. & Marquet, R. *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 (2002).
9. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127,** 4223–4231 (2005).
10. Mortimer, S. A. & Weeks, K. M. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.* **129,** 4144–4145 (2007).
11. Vasa, S. M., Guex, N., Wilkinson, K. A., Weeks, K. M. & Giddings, M. C. ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA* **14,** 1979–1990 (2008).
12. Gherghe, C. M., Shajani, Z., Wilkinson, K. A., Varani, G. & Weeks, K. M. Strong correlation between SHAPE chemistry and the generalized NMR order parameter (S^2) in RNA. *J. Am. Chem. Soc.* **130,** 12244–12245 (2008).
13. Pedersen, J. S., Meyer, I. M., Forsberg, R., Simmonds, P. & Hein, J. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res.* **32,** 4925–4936 (2004).
14. Leitner, T. *et al. HIV Sequence Compendium* (Theoretical Biology and Biophysics Group, 2005).
15. Purcell, D. F. & Martin, M. A. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *J. Virol.* **67,** 6365–6378 (1993).
16. Kwong, P. D. *et al.* Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **393,** 648–659 (1998).
17. Komar, A. A. A pause for thought along the co-translational folding pathway. *Trends Biochem. Sci.* **34,** 16–24 (2009).
18. Farabaugh, P. J. Programmed translational frameshifting. *Microbiol. Rev.* **60,** 103–134 (1996).
19. Wen, J. D. *et al.* Following translation by single ribosomes one codon at a time. *Nature* **452,** 598–603 (2008).
20. Nackley, A. G. *et al.* Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* **314,** 1930–1933 (2006).
21. Hartz, D., McPheeters, D. S., Traut, R. & Gold, L. Extension inhibition analysis of translation initiation complexes. *Methods Enzymol.* **164,** 419–425 (1988).
22. Mathews, D. H. *et al.* Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA* **101,** 7287–7292 (2004).
23. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure prediction. *Proc. Natl Acad. Sci. USA* **106,** 97–102 (2009).
24. Kim, D. H. *et al.* Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy. *Nature Biotechnol.* **23,** 222–226 (2005).
25. Stein, B. S. & Engleman, E. G. Intracellular processing of the gp160 HIV-1 envelope precursor. Endoproteolytic cleavage occurs in a cis or medial compartment of the Golgi complex. *J. Biol. Chem.* **265,** 2640–2649 (1990).
26. Wilson, W. *et al.* HIV expression strategies: ribosomal frameshifting is directed by a short sequence in both mammalian and yeast systems. *Cell* **55,** 1159–1169 (1988).
27. Giedroc, D. P., Theimer, C. A. & Nixon, P. L. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Mol. Biol.* **298,** 167–185 (2000).
28. Biswas, P., Jiang, X., Pacchia, A. L., Dougherty, J. P. & Peltz, S. W. The human immunodeficiency virus type 1 ribosomal frameshifting site is an invariant sequence determinant and an important target for antiviral therapy. *J. Virol.* **78,** 2082–2087 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This project was supported by the US National Institutes of Health (AI068462 to K.M.W.) and by the National Cancer Institute, under contracts N01-CO-12400 and HHSN261200800001E (to R.J.G. and J.W.B.). J.M.W. was supported as a Fellow of the UNC Lineberger Cancer Center and a National Institutes of Health (NIH) Kirschstein Postdoctoral Fellowship. R.S. and K.K.D. were supported by NIH grants AI44667 and T32 AI07419, respectively. We are indebted to D. Mathews and J. Low for assistance with the RNA structure program and genome secondary structure analysis, respectively. The content of this publication does not necessarily reflect the views or policies of the US Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations indicate endorsement by the US Government.

Author Contributions J.M.W., R.J.G. and K.M.W. conceived of and designed the HIV-1 genome structure analysis project. J.M.W. and K.M.W. analysed and interpreted the HIV SHAPE structure information. K.K.D., R.S. and C.L.B. designed and performed the bioinformatic pairing probability analysis. J.M.W., R.J.G. and C.W.L. performed the experiments. J.M.W., C.L.B. and K.M.W. performed the statistical analyses. J.W.B. produced and purified HIV-1 virions. J.M.W. and K.M.W. wrote the manuscript with contributions from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.M.W. (weeks@unc.edu).

METHODS

Virus production. HIV-1 strain NL4-3 (group M, subtype B) was used to infect a non-Hodgkin's T cell lymphoma cell line (a modified version of the SupT1 cell line, which was a gift from J. Hoxie)[29]. The virus-containing inoculum for infecting SupT1 cells was generated by $CaPO_4$/DNA coprecipitation[30] and subsequent transfection of pNL43 (NIH AIDS Research and Reference Reagent Program; GenBank accession AF324493) into 293T cells[31]. HIV-1 virions were purified as described[32] except cells were removed using a Millipore Opticap XL-5.0 micron filter. The total protein and CAp24 yields were 20.7 mg and 2.3 mg, on the basis of total protein (BioRad DC protein assay) and HPLC with subsequent amino acid analysis assays, respectively.

Virions were purified from cellular debris by subtilisin treatment and centrifugation through a sucrose cushion. Concentrated virions (in 19 ml, corresponding to 19 l of infected cell-culture supernatant) were digested with subtilisin (1 mg ml^{-1}, in 20 mM Tris (pH 8.0), 1 mM $CaC1_2$, 37 °C, 18 h; stopped by the addition of 5 μg ml^{-1} phenylmethylsulphonyl fluoride[33]). The resulting solution contained digested cellular proteins and viral particles free of surface proteins. The sample was centrifuged through a cushion of 20% (w/v) sucrose in PBS (Beckman SW41 rotor, 235,000*g*, 2 h, 4 °C); supernatant was carefully removed, and residual sucrose in the pellet was removed by overlaying PBS and repeating the centrifugation step (1 h at 4 °C).

RNA extraction. The key features of this protocol are that genomic RNA was gently extracted from purified virions in the presence of buffers containing monovalent and divalent ions, consistent with maintaining RNA secondary and tertiary structure. The HIV genomic RNA was not denatured by heat, chemical denaturants, magnesium chelation, or removal of monovalent cations during this process. Subtilisin-treated virions were suspended in virion lysis buffer (VLB; 50 mM HEPES (pH 8.0), 200 mM NaCl, and 3 mM $MgCl_2$) and lysed with 1% (w/v) SDS and 100 μg ml^{-1} proteinase K (~22 °C, 30 min). The digest was extracted three times with phenol/chloroform/isoamyl alcohol (25:24:1, pre-equilibrated with VLB), followed by two extractions with pure chloroform. Quantitative reverse-transcriptase PCR was used to quantify viral RNA yields against a standard curve[34–36]. The total yield from 19 l of infected cells was 97.2 pmol HIV-1 genomic RNA. The aqueous layer (3.6 ml) was brought to 300 mM NaCl and precipitated with 70% (v/v) ethanol. Retroviral genomes commonly contain single-stranded breaks[2]. Approximately 30% of our genomic RNA was intact, as judged by visualization in agarose/formaldehyde gels; nicks in the remaining 70% seemed to be roughly randomly distributed on the basis of direct visualization of the genomic RNA and from the continuity of our primer extension reactions (see Supplementary Table 1).

RNA modification. The RNA pellet containing 97.2 pmols of HIV-1 genomic RNA was dissolved in 880 μl of modification buffer (50 mM HEPES (pH 8.0), 200 mM potassium acetate (pH 8.0), 3 mM $MgCl_2$) and incubated at 37 °C for 15 min. Then, 405 μl of the solution was added to 45 μl pre-warmed (37 °C) 1M7 (in dimethylsulphoxide (DMSO))[10] or to DMSO. After 4 min, 45 μl of 50 mM EDTA (pH 8.0) were added to each tube. The reactions were divided into 11-μl aliquots and precipitated with ethanol.

Primer synthesis. Primers were designed with the aid of OligoWalk, part of the RNAstructure software package[22] (available for download at http://rna.urmc.rochester.edu/) (Supplementary Table 1). Primers were required to be 20–22 nucleotides in length, have high melting temperatures and low self-annealing energies, and preferably end with a 3′ G or C. Only 2 out of 31 primers required redesign, giving OligoWalk a 94% success rate. Primers were synthesized to contain a 5′ six carbon linker terminating in a primary amine (IDT). The amine-tethered DNA primers (1 μl; 25 μg ml^{-1}) were labelled with one of four fluorophores (5-FAM, 6-JOE, 6-TAMARA or 5-ROX; AnaSpec) using *N*-hydroxysuccinimide chemistry (3 μl NHS-coupled dye (20 mg ml^{-1} in DMSO) in 0.1 M $NaBO_3$-HCl (pH 8.5); ~22 °C, 3 h). Labelled primers were precipitated with ethanol, purified on a denaturing gel (20% 29:1 acrylamide/bis-acrylamide, 7 M urea, 1× TBE), recovered by passive elution in water, precipitated (300 mM NaCl, 2.5 vol ethanol, 1 vol isopropanol), pelleted, and dissolved in water. Spectrophotometric measurements indicated labelling was ~90–95% efficient as determined by the [dye]/[DNA] ratio.

Primer extension. RNA pellets (1 pmol) were dissolved in 10 μl 0.5× TE (5 mM Tris (pH 8.0), 0.5 mM EDTA) and mixed with 3.0 μl of 0.4 μM primer. The (+) and (−) 1M7 reagent reactions were labelled with JOE and FAM, respectively. Primers were annealed to the RNA by heating to 65 °C for 5 min and 45 °C for 2 min, and then placed on ice. Six microlitres of reverse transcriptase mix[37] (SuperScript III, 5× buffer, DTT, dNTPs; Invitrogen) was added to each tube and incubated for 10 s at 45 °C, 5 min at 52 °C, 5 min at 65 °C, and cooled to 4 °C. Sodium acetate (pH 5.2; 2.0 μl at 3 M) was added to each tube, (+) and (−) 1M7 tubes were combined, and 120 μl of ethanol was added to precipitate the cDNA products. The reactions were pelleted, washed with 70% ethanol, and dissolved in 10 μl deionized formamide.

Sequencing. Dideoxy sequencing reactions (GenomeLab Methods Development Kit; Beckman) were performed using plasmids pDR0 and pDR25 (containing partial NL4-3 sequences), and primers were labelled with TAMARA and ROX. Primer sequences were identical to those in Supplementary Table 1 except primer 31, the sequence (5′-CTGCAACCTCTACCTCCTG GGTGCTAGAG-3′) of which annealed to the plasmid rather than the poly(A) RNA sequence in the genomic RNA.

Capillary electrophoresis. cDNA fragments were resolved by capillary electrophoresis[6,10] (Applied Biosystems AB3130 instrument). Samples were injected at 1.2 kV for 16 s into a 36-cm capillary containing POP7 (ABI) and subjected to electrophoresis for 25 min at 15 kV. The fluorescence detector was initially calibrated with 5-FAM, 6-JOE, 6-TAMARA and 5-ROX using fluorescent markers with fragment lengths of 242 (5-FAM), 206 (6-JOE), 188 (6-TAMARA) and 155 (5-ROX) nucleotides. Fragments were generated by linear amplification of HindIII-digested plasmid pUC18 using primers with the sequences 5′-CAGAGCAGATTGTACTGAGAG-3′, 5′-GTGAAATACCGCAC AGATGC-3′, 5′-GCGTAAGGAGAAAATACCGCATC-3′ and 5′-CGCCATTC AGGCTGCGCAACTG-3′, labelled with 5-FAM, 6-JOE, 6-TAMARA and 5-ROX, respectively. Fluorescent spectral overlap based on this DNA ladder was calibrated using AB3130 software.

Data processing. Raw electropherograms, containing fluorescence intensity versus elution time information, were converted to normalized SHAPE reactivities using ShapeFinder[6,11,23] (available for download at http://bioinfo.unc.edu). The ShapeFinder software aligns the (+) and (−) reagent traces to the two dideoxy nucleotide sequencing ladders, corrects for signal decay[38], and performs a whole-channel Gaussian integration[11] to quantify all individual peak areas (see Fig. 3a). Only 11 of the 9,173 nucleotides in the NL4-3 RNA genome had high background and were therefore excluded from analysis. Data sets were normalized to a scale such that 1.0 represents the average intensity of highly reactive nucleotide positions[6,23]. On this scale, ~95% of integrated intensities for the HIV-1 genome fall between 0 and 1 (see histogram in Fig. 3b). Each primer extension reaction was processed individually. The resulting intensities in regions with overlapping data from different primers correlated closely: reactivity differences were typically less than 0.1 SHAPE unit. Regions with overlapping data accounted for ~25% of the total nucleotide positions and were averaged to generate the final data set spanning the entire NL4-3 genome.

Toeprinting ribosome pause sites at the matrix–capsid and capsid–nucleocapsid junctions. A double-stranded DNA template to direct synthesis of an mRNA spanning NL4-3 Gag nucleotides 1 to 1795 was generated by PCR. This region encompasses the entire 5′ UTR and most of the *gag* coding region and ends after the three-stem frameshift element. The 5′ primer included a T7 promoter sequence (5′-TAATACGACTCACTAATGGTCTCTCTGGTTAGACCA-3′), and the 3′ primer (5′-GCTAAAGGTTACAGTTCCTTGTC-3′) encoded a stop codon at position 1787. The RNA transcript was capped and polyadenylated (mSCRIPT, Epicentre) and *in vitro* translation was carried out in rabbit reticulocyte extract (Ambion) using ~60 μg of the capped, polyadenylated transcript, 1 μl 1.25 mM L-methionine, 1 μl ^{35}S-methionine (PerkinElmer), 17 μl reticulocyte extract, and 1.25 μl 20× 'medium-salt' translation buffer (Ambion) in a total volume of 26 μl at 37 °C. Cycloheximide was added at 0, 7, 15 or 45 min to arrest translation[21]. Translation reaction aliquots were separated on an 8–16% SDS–PAGE gel (Invitrogen) to confirm production of a protein of the correct length. Sites of ribosome pausing were detected by adding the following to 25 μl of the *in vitro* translation mixture: 1.35 μl 10 mM each dNTP, 2 μl 4.0 μM fluorescently labelled primer (primer 4 or 6 for interrogating the matrix–capsid and capsid–nucleocapsid regions, respectively), 1 μl 200 mM $MgCl_2$, and 2 μl Superscript III (Invitrogen). The translation reaction that was pre-quenched with cycloheximide was taken as background and was resolved using a JOE-labelled primer. The 7, 15 and 45 min time points were resolved using FAM-labelled primers. Primer extension reactions were incubated at 37 °C for 30 min and stopped by the addition of 1 μl 0.5 M EDTA and 400 μl water. The reaction was extracted with phenol:chloroform:isoamyl alcohol (25:24:1, 2×) and chloroform (1×). Four microlitres of this solution, 1 μl of a cDNA sequencing ladder, and 15 μl of formamide were combined, heated to 105 °C for 5 min, and resolved by capillary electrophoresis. Toeprinting traces were processed with ShapeFinder[11] and normalized to a scale in which 1.0 is equal to the mean intensity of the most reactive positions, identically as described above for SHAPE experiments.

RNA secondary structure model. The entire NL4-3 sequence—9,173 nucleotides plus 20 3′ adenosines (representing the poly(A) tail)—was folded using the thermodynamics-based algorithm in RNAstructure[22,23]. SHAPE information was used to constrain secondary structure calculations by incorporating SHAPE reactivities as pseudo free-energy change terms[6,23] using slope and intercept values of 30 and −6, respectively. The maximum distance allowed between any two paired positions was 600 nucleotides. The slope and intercept values are derived from previous parameterization on long RNAs, and the 600-nucleotide

cutoff reflects that 99% of all base pairs in ribosomal RNA occur between nucleotides less than this distance apart[23]. The genome was initially folded as a single (9,193 nucleotides) unit; folding was then fine-tuned by dividing the RNA into five independent folding regions, separated by long stretches of reactive nucleotides that were calculated to be unpaired when the entire genome was folded with SHAPE constraints (NL4-3 residues 1–2844, 2836–5722, 5676–6832, 6807–7791 and 7779–9193). Dividing the genome in this way facilitated model building and prevented the formation of a few poorly supported long-distance pairings between domains. Highly reactive nucleotides at the termini of each region were prohibited from forming base pairs in these region-specific calculations. Helices consisting of a single base pair were removed from the final model and unreactive nucleotides in the primer binding site (183–199) were taken to reflect hybridization with the tRNA primer. The current version of our algorithm does not allow pseudoknots and therefore our HIV-1 structure model (Fig. 2) includes only one, heuristically predicted[6,8], pseudoknot.

Quality of SHAPE-directed model of the entire HIV-1 genome. The algorithm by which SHAPE information is used to create an RNA secondary structure model does not make any specific assumptions about the magnitude of SHAPE reactivity that corresponds to single-stranded versus based-paired regions. Instead, SHAPE reactivities are converted to free-energy change terms and used to constrain a thermodynamic folding algorithm[22,23]. SHAPE information is essential for generating this secondary structure model. Folding the genome by free-energy minimization alone, using a best-of-class algorithm[22,39], results in a structure that is very different from the experimentally supported model. Only 47% of the base pairs in the SHAPE-directed model also occur in the lowest free-energy thermodynamics-only model. The unconstrained thermodynamics-only model is readily shown to be incorrect because many regions with high SHAPE reactivities are assigned as paired in the unconstrained model, and many regions with low SHAPE reactivities are assigned as single-stranded.

Several lines of evidence support fundamental correctness of our working SHAPE-directed HIV-1 genome structural model (Fig. 2). First, SHAPE-directed folding is well validated and predicts the known structures of large RNAs, including 16S ribosomal RNA, with high accuracies (>90%)[10,23]. Second, most nucleotides assigned to single-stranded regions are reactive by SHAPE (Fig. 2, red, orange, and green nucleotides). Conversely, base-paired nucleotides are generally unreactive (Fig. 2, black nucleotides and inset). Thus, the structural modelling faithfully incorporates the experimental data. Third, many single nucleotide bulges are predicted as single reactive positions imbedded in helices with flanking nucleotides that are unreactive towards SHAPE, which speaks to the accuracy at the single nucleotide resolution level (for select examples see Fig. 2, positions 1725, 3376, 4891, 5990, 7431, 7568 and 9091). Fourth, previously characterized HIV RNA structures including the 5′ TAR element, the DIS component of the packaging signal, and the five-stem RRE, serve as positive controls and form structures consistent with previous work[4,40] (Fig. 2). In the case of the *gag-pol* frameshift structure, we note that SHAPE data do not support common alternative proposals for this specific structure, including either a longer bulged stem or a pseudoknot.

Most structures in our current HIV-1 genome model, especially in regions with several closely spaced helices, are extremely well determined, as evidenced by the strong correlation between SHAPE values and base pairing. This correlation is also consistent with benchmarking studies showing that SHAPE reactivities strongly discriminate between base paired and single-stranded nucleotides (Supplementary Fig. 2)[41] and are proportional to the extent of local nucleotide disorder[12]. In contrast, some of the larger loop regions in our model may reflect regions that interconvert between multiple structures[38,42]. Elements that may require future refinement include the precise termini of helices at some multi-helix junctions and along the central backbone of the genome structure and the identification of further pseudoknot and long-range interactions.

Calculation of evolutionary base pairing probabilities. RNA-Decoder[13] was used to identify regions in the HIV-1 genome in which the ability to form base pairs is evolutionarily conserved. The program takes a set of grammar parameters, a multiple-sequence alignment, and a phylogenetic tree as input. The output is a pairing probability for each position in the genome, given the phylogenetic tree, alignment, and the grammar structural model. The pairing probability for position i in alignment D is the sum over all stem structural labels (k) of $P(\pi_i = k|M)P(D|\pi,T,M)$, where π is the structure, M is the grammar model parameters, and $P(\pi_i = k|M)$is the posterior probability that position i has the specific structural label k, given the grammar[43], and is calculated by the inside-outside algorithm[44]. In Bayesian terms, $P(\pi_i = k|M)$ is the prior probability of structure π and $P(D|\pi,T,M)$ is the alignment probability, calculated using the Felsenstein algorithm[45]. Pairing predictions were made using an alignment of non-recombinant group M subtype reference sequences obtained from the Los Alamos HIV database[14], with minor manual editing (and excluding subtype G, which is now considered a circulating recombinant form[46]). Codon positions in genome regions encoding more than one protein in overlapping reading frames were defined according to the first open reading frame in the following pairs: *gag-pro*, *pol-vif*, *vpr-vif*, *vpr-tat*, *rev-tat*, *env-vpu*, *env-tat2* and *env-rev2*. Owing to differences in nucleotide content and evolution rate within different genes in the HIV genome, the genome was scanned in two sections, upstream and downstream, that overlapped in the *vif* gene. This allowed use of separate phylogenetic trees for each scan, with branch lengths calculated according to the rates of evolution in each genome region. The phylogenetic tree for the 5′ half was built using the third codon position for the *gag*, *pol* and *vif* genes, and the 5′ non-coding region; the tree for the 3′ half was built on the third positions of *vif*, *vpr*, *rev*, *vpu*, *env* and *nef* genes, and the 3′ non-coding region.

Pairing probabilities were assessed across the entire genome. To accommodate as many pairing interactions as possible, we used a large window size (1,300 nucleotides), and spaced the scans at 300-nucleotide intervals. Pairing probabilities for each scan were combined using the statistical program R[47] taking the maximum pairing probability in overlapping windows. It is important to note that high pairing probabilities identify regions experiencing evolutionary pressure to retain a specific, defined, secondary structure. A low pairing probability, although suggestive of a lack of structure, can also reflect (1) that an additional evolutionary constraint exists that is not accounted for by the evolutionary model, or (2) that natural selection favours folding in general, but not a precise pattern of folding.

Bootstrap analysis of SHAPE reactivities in inter-protein linkers and protein-domain junction. A bootstrap procedure was used to compare the SHAPE reactivities of particular collections of genome elements to the expectation for random genome regions of the same size. For a comparison to a collection of n genome elements, we generated 100,000 bootstrapped samples by randomly choosing n locations from the relevant portion of the genome, and randomly assigning the lengths of the actual genome elements to these n locations. For comparison to the protein-domain junctions, locations were drawn randomly from the entire coding portion of the genome (bases 336–8621). We specified a length of 60 nucleotides for each region. For comparison to the intra-domain loops, locations were drawn randomly from within the domains where loops occur and assigned lengths that reflected loop sizes in the same domain (for example, for the capsid domain, one element of 45 base pairs was drawn from within bases 732–1427). Bootstrap samples that contained overlapping genome regions were thrown out. The mean SHAPE reactivities for each bootstrap sample were used to generate a frequency distribution that describes the expectation for equally sized but randomly located collections of genome elements in HIV coding regions. We obtained a P value by determining the percentage of the bootstrapped means that was lower than the mean SHAPE reactivity for the collection of genome elements. This P value is equivalent to the probability that the low SHAPE reactivity in the actual collection of genome elements occurred by chance. P values for inter-protein linkers and protein-domain junctions were 0.0482 and 0.0777, respectively. The reverse transcriptase–RNase H junction functions both as an inter-protein linker and as a protein-domain junction because it is cleaved one-half of the time. For this analysis, the reverse transcriptase–RNase H junction was counted as an inter-protein linker.

Statistical analysis of ribosome pause sites. Toeprinting data spanned 748 nucleotides (positions 670–1018 and 1243–1652; Supplementary Fig. 5). In these two reads, there were 220 nucleotides that fell within 30 nucleotides of the matrix–capsid, capsid–nucleocapsid, or nucleocapsid–p6 junctions or in the cyclophilin loop. We evaluated whether ribosomes pause preferentially near protein junctions using the binomial distribution. A total of 36 base pairs yielded toeprint signals with an intensity of 1.0 or greater. A signal of 1.0 corresponds approximately to 1.5 standard deviations above the mean; 17 of these occurred within 30 nucleotides of a protein junction. The probability of observing this distribution by chance is $P = 0.018$. This analysis was insensitive to the choice of high signal threshold. Similar P values were obtained for toeprint thresholds between 0.6 and 1.6.

Consensus structure. The *gag-pro-pol* consensus structure (Supplementary Fig. 6b) was generated by aligning the 37 reference group M HIV-1 sequences[14] using CLUSTALW[48]. Regions of covariation were identified using a sequence logo[49].

Helix energies. Helix free-energy changes (Fig. 4c, d) were calculated using the RNAstructure program[22] as the sum of the base pair stacking nearest neighbour parameters[50,51]. Duplex regions containing single nucleotide bulges were taken to be a single helix. The helix free-energy changes do not include penalties for terminal AU or GU pairs because these are, by convention in RNAstructure, associated with the loop formation free-energy changes.

RNA and protein structure display. RNA secondary structures were composed using xrna (http://rna.ucsc.edu/rnacenter/xrna); HIV protein images (Fig. 1d) were created using Visual Molecular Dynamics[52].

29. Means, R. E. *et al.* Ability of the V3 loop of simian immunodeficiency virus to serve as a target for antibody-mediated neutralization: correlation of neutralization

sensitivity, growth in macrophages, and decreased dependence on CD4. *J. Virol.* **75,** 3903–3915 (2001).
30. Graham, F. L. & van der Eb, A. J. A new technique for the assay of infectivity of human adenovirus 5 DNA. *Virology* **52,** 456–467 (1973).
31. Adachi, A. *et al.* Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. *J. Virol.* **59,** 284–291 (1986).
32. Chertova, E. *et al.* Envelope glycoprotein incorporation, not shedding of surface envelope glycoprotein (gp120/SU), is the primary determinant of SU content of purified human immunodeficiency virus type 1 and simian immunodeficiency virus. *J. Virol.* **76,** 5315–5325 (2002).
33. Ott, D. E. *et al.* Analysis and localization of cyclophilin A found in the virions of human immunodeficiency virus type 1 MN strain. *AIDS Res. Hum. Retroviruses* **11,** 1003–1006 (1995).
34. Thomas, J. A. *et al.* Human immunodeficiency virus type 1 nucleocapsid zinc-finger mutations cause defects in reverse transcription and integration. *Virology* **353,** 41–51 (2006).
35. Cline, A. N., Bess, J. W., Piatak, M. Jr & Lifson, J. D. Highly sensitive SIV plasma viral load assay: practical considerations, realistic performance expectations, and application to reverse engineering of vaccines for AIDS. *J. Med. Primatol.* **34,** 303–312 (2005).
36. Buckman, J. S., Bosche, W. J. & Gorelick, R. J. Human immunodeficiency virus type 1 nucleocapsid Zn^{2+} fingers are required for efficient reverse transcription, initial integration processes, and protection of newly synthesized viral DNA. *J. Virol.* **77,** 1469–1480 (2003).
37. Wilkinson, K. A., Merino, E. J. & Weeks, K. M. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols* **1,** 1610–1616 (2006).
38. Badorrek, C. S. & Weeks, K. M. Architecture of a gamma retroviral genomic RNA dimer. *Biochemistry* **45,** 12664–12672 (2006).
39. Dowell, R. D. & Eddy, S. R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* **5,** 71 (2004).
40. Olsen, H. S., Nelbock, P., Cochrane, A. W. & Rosen, C. A. Secondary structure is the major determinant for interaction of HIV rev protein with RNA. *Science* **247,** 845–848 (1990).
41. Wilkinson, K. A. *et al.* Influence of nucleotide identity on ribose 2′-hydroxyl reactivity in RNA. *RNA* **15,** 1314–1321 (2009).
42. Badorrek, C. S. & Weeks, K. M. RNA flexibility in the dimerization domain of a gamma retrovirus. *Nature Chem. Biol.* **1,** 104–111 (2005).
43. Knudsen, B. & Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.* **31,** 3423–3428 (2003).
44. Durbin, R. & Eddy, S. *Biological Sequence Analysis: Probabalistic Models Of Proteins And Nucleic Acids* 356 (Cambridge Univ. Press, 1998).
45. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17,** 368–376 (1981).
46. Abecasis, A. B. *et al.* Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. *J. Virol.* **81,** 8543–8551 (2007).
47. The R Development Core Team. *The R Foundation for Statistical Computing* <http://www.R-project.org> (2008).
48. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
49. Chang, T. H., Horng, J. T. & Huang, H. D. RNALogo: a new approach to display structural RNA alignment. *Nucleic Acids Res.* **36,** W91–W96 (2008).
50. Xia, T. *et al.* Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs. *Biochemistry* **37,** 14719–14735 (1998).
51. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. *J. Mol. Biol.* **288,** 911–940 (1999).
52. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* 14, 33–38 27–38 (1996).