Article

Determination of RNA structural diversity and its role in HIV-1 RNA splicing

https://doi.org/10.1038/s41586-020-2253-5

Received: 12 May 2019

Accepted: 4 March 2020

Published online: 06 May 2020

Check for updates

Phillip J. Tomezsko[1,2,3,18], Vincent D. A. Corbin[4,5,18], Paromita Gupta[1,18], Harish Swaminathan[1], Margalit Glasgow[1,6], Sitara Persad[1,6], Matthew D. Edwards[7], Lachlan Mcintosh[4,8,9], Anthony T. Papenfuss[4,5,8,9,10], Ann Emery[11,12,13], Ronald Swanstrom[12,13,14], Trinity Zang[15], Tammy C. T. Lan[1], Paul Bieniasz[15,16], Daniel R. Kuritzkes[3,17], Athe Tsibris[3,17] & Silvi Rouskin[1]✉

Human immunodeficiency virus 1 (HIV-1) is a retrovirus with a ten-kilobase single-stranded RNA genome. HIV-1 must express all of its gene products from a single primary transcript, which undergoes alternative splicing to produce diverse protein products that include structural proteins and regulatory factors[1,2]. Despite the critical role of alternative splicing, the mechanisms that drive the choice of splice site are poorly understood. Synonymous RNA mutations that lead to severe defects in splicing and viral replication indicate the presence of unknown *cis*-regulatory elements[3]. Here we use dimethyl sulfate mutational profiling with sequencing (DMS-MaPseq) to investigate the structure of HIV-1 RNA in cells, and develop an algorithm that we name 'detection of RNA folding ensembles using expectation–maximization' (DREEM), which reveals the alternative conformations that are assumed by the same RNA sequence. Contrary to previous models that have analysed population averages[4], our results reveal heterogeneous regions of RNA structure across the entire HIV-1 genome. In addition to confirming that in vitro characterized[5] alternative structures for the HIV-1 Rev responsive element also exist in cells, we discover alternative conformations at critical splice sites that influence the ratio of transcript isoforms. Our simultaneous measurement of splicing and intracellular RNA structure provides evidence for the long-standing hypothesis[6–8] that heterogeneity in RNA conformation regulates splice-site use and viral gene expression.

Previous work[4] on the genome-wide RNA structure of HIV-1 in vitro and in virion has provided a population-average model, with the underlying assumption that every molecule within the population assumes the same conformation. However, previous in vitro studies[5,9,10] have identified alternative conformations for the HIV-1 Rev responsive element (RRE) and 5′ untranslated region (UTR), which raises the possibility that alternative structures have roles in the export of viral RNA from the nucleus and packaging in virions. To resolve the fundamental question of whether RNA structure affects splicing, it is necessary to distinguish multiple conformations for the same sequence in cells. We developed the clustering algorithm DREEM, and here we demonstrate that we can quantitatively detect alternative RNA structures.

DREEM starts with single-molecule chemical probing data—in our case, derived from DMS-MaPseq[11]. Dimethyl sulfate (DMS) adds methyl groups to the unpaired adenines and cytosines of RNA molecules (Fig. 1). The presence of a methyl adduct is read during reverse transcription using TGIRT-III, which marks these sites by incorporating random mutations in the complementary (c)DNA. PCR amplifies the cDNA product and attaches sequencing adapters to the DNA, followed by massively parallel sequencing. Each of the resulting reads is represented as a binary readout of mutations and matches, which is the input for DREEM (Extended Data Fig. 1a). As DMS-MaPseq has a negligible background error[11], the mutations observed on a single DNA molecule correspond to the DMS-accessible bases on the parent RNA molecule. The two key challenges for detecting heterogeneity are: (1) that DMS modification rates are relatively low (for example, an open base has a probability of about 2–10% of being modified); and (2) that the rate of DMS modification per open base is sensitive to the local chemical environment, such that not all open bases are equally reactive to DMS. Traditional approaches to determining RNA structure combine chemical probing data into a population-average signal per base, which obscures any underlying heterogeneity. By contrast, the

[1]Whitehead Institute for Biomedical Research, Cambridge, MA, USA. [2]Program in Virology, Harvard Medical School, Boston, MA, USA. [3]Brigham and Women's Hospital, Boston, MA, USA. [4]Bioinformatics Division, Walter and Eliza Hall Institute, Parkville, Victoria, Australia. [5]Department of Medical Biology, The University of Melbourne, Melbourne, Victoria, Australia. [6]Massachusetts Institute of Technology, Cambridge, MA, USA. [7]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [9]Department of Mathematics and Statistics, University of Melbourne, Melbourne, Victoria, Australia. [10]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, Victoria, Australia. [11]Curriculum in Genetics and Molecular Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [12]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [13]Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [14]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [15]Laboratory of Retrovirology, The Rockefeller University, New York, NY, USA. [16]Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA. [17]Department of Medicine, Harvard Medical School, Boston, MA, USA. [18]These authors contributed equally: Phillip J. Tomezsko, Vincent D. A. Corbin, Paromita Gupta. ✉e-mail: srouskin@wi.mit.edu

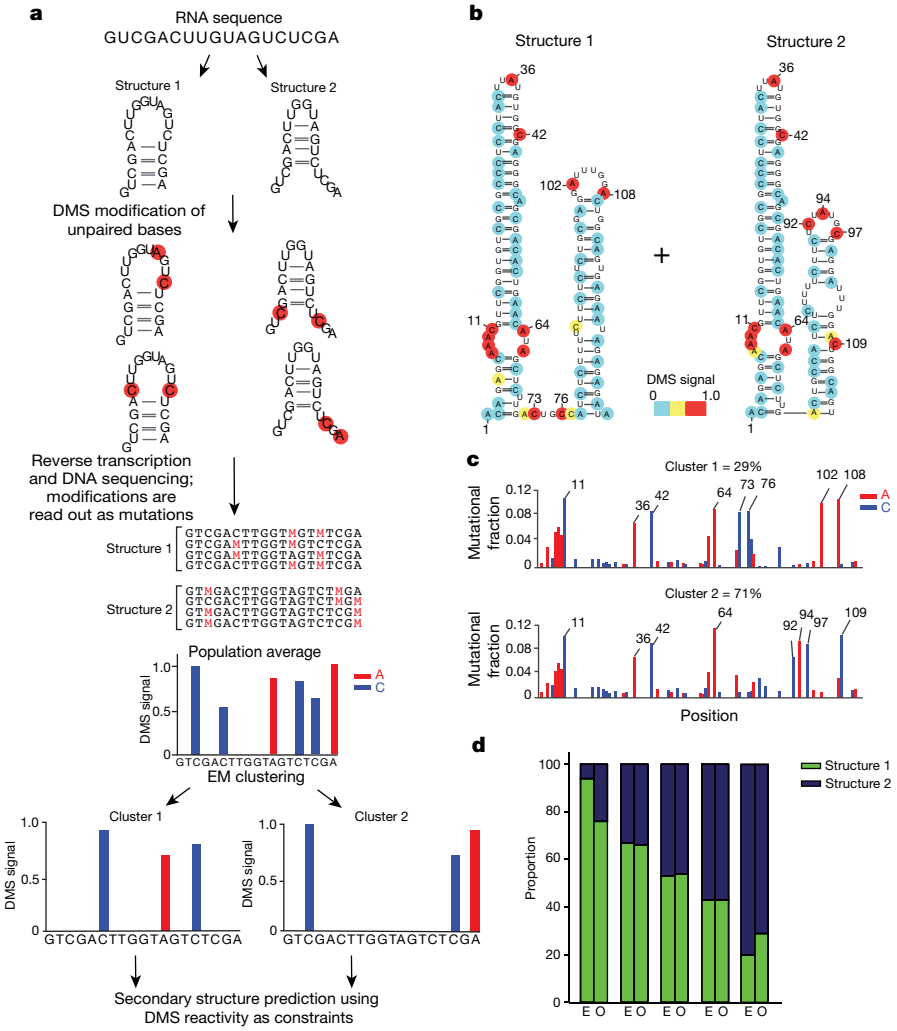


Fig. 1 | Development and validation of DREEM algorithm for analysis of alternative RNA structures. a, Schematic of combining DMS-MaPseq data with the DREEM algorithm to detect alternative RNA structures. EM, expectation–maximization. **b**, Structural model of in vitro-transcribed and folded structure 1 and structure 2, as determined by DMS-MaPseq. Nucleotides are colour-coded by normalized DMS signal. **c**, DMS mutational fraction per nucleotide and quantification of structure 1 and structure 2, determined by DREEM clustering for a mixing ratio of 25% (structure 1) to 75% (structure 2) before DMS modification. **d**, Proportion of structure 1 and structure 2 measured by DREEM clustering after in vitro transcription, mixing and DMS-MaPseq. The expected (E) and observed (O) ratios are shown from $n = 1$ experiment for each mixing proportion.

DREEM algorithm groups sequencing reads issued from each structure into distinct clusters, by exploiting information contained in the observation of multiple modifications on single molecules. Theoretically, if two individual bases are DMS-reactive in the population average but are never both mutated on a single read, it follows that at least two conformations are present. DREEM identifies patterns of DMS-induced mutations on reads and clusters in a mathematically rigorous manner using an expectation–maximization algorithm (Fig. 1a, Extended Data Fig. 1a). The DMS modification rate per base for each cluster (or structure) is determined by iteratively maximizing a log-likelihood function to find and quantify the abundance of alternative structures directly from the dataset. The binary nature of the readouts enables the use of a multivariate Bernoulli mixture model to compute the log-likelihood function[12]. The DMS modification pattern from each cluster is used to create a secondary structure model.

Our control experiments on denatured RNA indicated that TGIRT-III is unable to read-through mismatches located within three nucleotides (nt) of each other (Extended Data Fig. 1b). To account for this observation, we modified the log-likelihood function of the standard multivariate Bernoulli mixture model (Extended Data Fig. 1a). Upon convergence of the clustering, the DMS signal from each cluster was used as a constraint in the program RNAstructure[13]. To our knowledge, DREEM is unique among algorithms for RNA folding ensembles[14] because DREEM directly clusters the experimental data. Clustering before the generation of a secondary structure model enables the discovery of new RNA structures, in contrast to previous work[15,16]. Purely computational algorithms rely on suboptimal folding to create variation that is not captured by minimum free energy calculations. However, using experimentally derived constraints is preferable to using randomly generated constraints[17,18]. Moreover, DREEM does not rely on thermodynamics for detecting and identifying alternative conformations, and therefore can be used on in vivo data to model RNA folding in the presence of cellular factors, the energetic contributions to RNA structure of which are unknown.

To validate DREEM, we first transcribed two RNA molecules in vitro that are nearly identical in sequence but form different structures (which we refer to as structure 1 and structure 2). These sequences were designed on the basis of a known RNA structure that is changed by a single nucleotide variant (riboSNitch) in the human gene *MRPS21*[19]. We experimentally mixed the RNAs from both structures in varying proportions and generated DMS-MaPseq data. DREEM clustered the DMS data and successfully identified the two structures, down to a

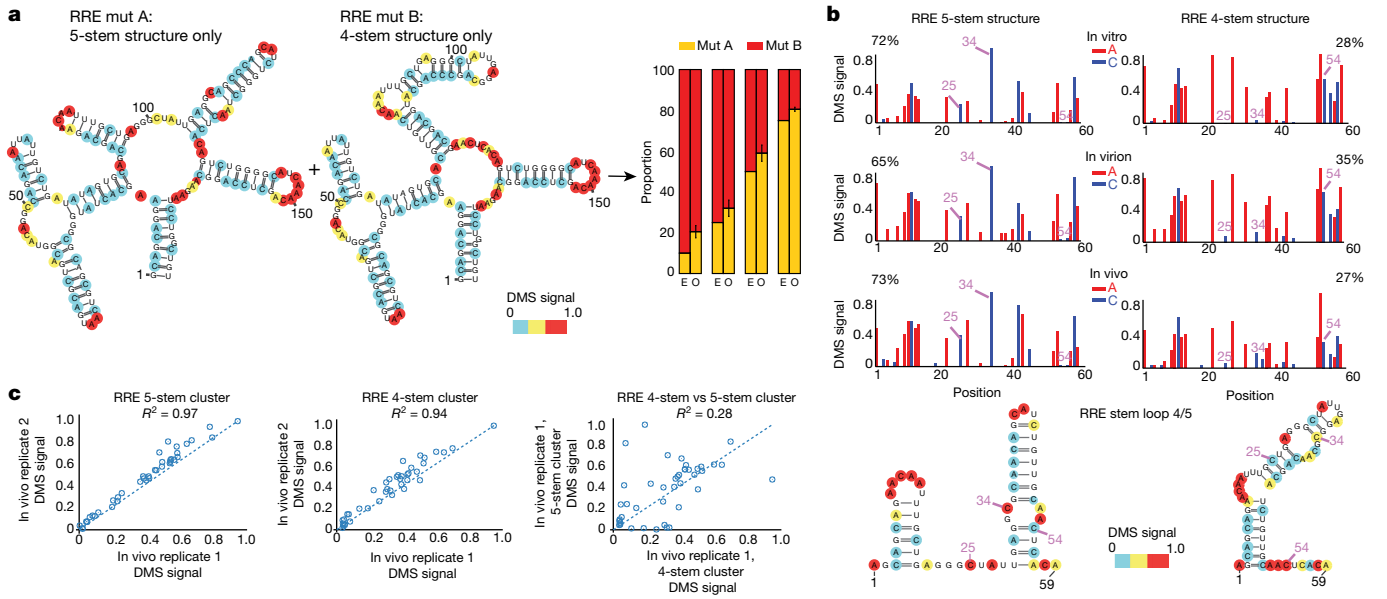


Fig. 2 | The formation of alternative structures at HIV-1 RRE is driven by intrinsic RNA thermodynamics. a, HIV-1 RRE structural models derived from DMS-MaPseq followed by DREEM using in vitro-transcribed structure-locked RRE five-stem (mut A) and four-stem (mut B) mutants. Bar graphs represent expected and observed mixing ratios of four-stem and five-stem structures from $n = 2$ experiments. **b**, Normalized DMS signal for RRE five-stem and four-stem structures observed in vitro, in virion and in vivo from CD4$^+$ T cells infected with HIV-1_{NL4-3}, identified by DREEM clustering. The positions highlighted are examples of bases that change pairing state between the two structures, shown in both the DMS signal and the folded RNA structures of the four-stem and five-stem structures. Percentages for each cluster are determined by DREEM from representative samples of $n = 2$ (for in vivo and in vitro), or from $n = 1$ for in virion. **c**, Scatter plots of clustering results for $n = 2$ biological replicates (top two plots) and the variation in DMS signal between the different two clusters (four-stem versus five-stem, bottom).

mixing ratio of 6% (Fig. 1b–d, Extended Data Fig. 2). We also tested DREEM using the in vitro-transcribed and DMS-modified adenosine deaminase (*add*) riboswitch, which undergoes a conformational shift upon binding of adenine[20,21]. We found that *add* structures that promote translation, which are stabilized by adenine, increased from 18% to 89% of the structures detected by DREEM upon addition of 5 mM adenine (Extended Data Fig. 3).

We then focused on the RRE of HIV-1_{NL4-3}, a multi-stem structure that binds to the viral protein Rev and enables the nuclear export of unspliced and partially spliced HIV-1 RNA. Previous studies[5] physically separated distinct RNA conformations by native gel electrophoresis, and revealed two alternative structures for RRE in vitro: a five-stem and a four-stem structure. Specific mutations stabilize either of the alternative conformations[5]. DREEM accurately identified the DMS signal for mixtures of five-stem structures (referred to here as mut A) and four-stem structures (referred to here as mut B), and robustly quantified their mixing ratios (Fig. 2a). Notably, we found that the in vitro-folded wild-type RRE sequence exists as a mixture of about 27% four-stem and about 73% five-stem structures (Extended Data Fig. 4).

We next applied DREEM to the study of HIV-1 RNA structure in primary cells, which is possible as DMS is cell-membrane permeable[22]. We infected activated CD4$^+$ T cells with HIV-1_{NL4-3}, and performed chemical probing in vivo and in virions (Extended Data Fig. 5a). We discovered that the RRE sequence forms the same alternative structures regardless of the environment (in vitro, in vivo or in virion), favouring the five-stem fold (Fig. 2b). These results indicate that the alternative secondary structures of RRE are driven largely by intrinsic RNA thermodynamics as opposed to particular features of the cellular environment. Moreover, these results underscore the ability of DREEM to robustly identify RNA folding ensembles from in vivo data (Fig. 2b, c, Extended Data Fig. 5b) and to quantify the abundance of the alternate conformations.

We next examined the role of RNA structure in HIV-1 splicing. Alternative splicing is the major mechanism that is used by HIV-1 to express all of its gene products from a single type of pre-mRNA (that is, genomic viral RNA). Splice site use must be regulated to produce the correct proportion of HIV-1 transcripts. HIV-1 transcripts spliced at the A3 acceptor splice site are the only source of mRNA for the viral transcriptional activator Tat[1,2].

We discovered alternative structures that dictate the splicing outcome at the A3 splice site, and therefore regulate the abundance of the Tat transcript. First, the structures that form for the HIV-1_{NL4-3} A3 splice site in CD4$^+$ T cells differ markedly from previously proposed models based on population-average data[4]. Notably, the two main conformations identified by DREEM either occlude (about 40%; cluster 1) or expose (about 60%; cluster 2) the polypyrimidine tract and A3 splice site at which the U2AF heterodimer binds (here we abbreviate the tract and splice site together as A3ss) (Fig. 3a). We termed the occluded structure the A3 stem loop (A3SL). The A3SL is not specific to the HIV-1_{NL4-3} and forms in HIV-1_{NHG} in HEK293T cells (cluster 1 in Fig. 3b). Notably, we detected strong heterogeneity for the A3ss folded in vitro, demonstrating that this region has an intrinsic ability to form multiple conformations and that the A3SL is thermodynamically stable in the absence of proteins (Extended Data Fig. 6).

To perturb the population of RNA structures and measure the effect on splicing, we took advantage of the location of A3ss in the *vpr* coding region, which is dispensable for growth in cell culture. We used a strain (Δ*vpr* HIV-1_{NHG}) with a pre-mature stop codon in *vpr* to ensure that the observed effects were not due to loss of function of Vpr. To test the effect of structure on splicing, we designed mutations distal from the splice site sequence, avoiding known protein-binding regions. The mutants A3SL mut1, mut 2 and mut 3 are predicted to thermodynamically stabilize A3SL and decrease splicing at A3ss (Fig. 3c). Using a deep-sequencing-based HIV-1 splicing assay[23], we found that all three stabilizing mutants result in lower rate use of A3ss (Fig. 3d), substantially decreasing expression of Tat transcripts relative to a background strain.

By contrast, mutations in the same sequence region that are predicted to have little effect on the stability of A3SL—and therefore little

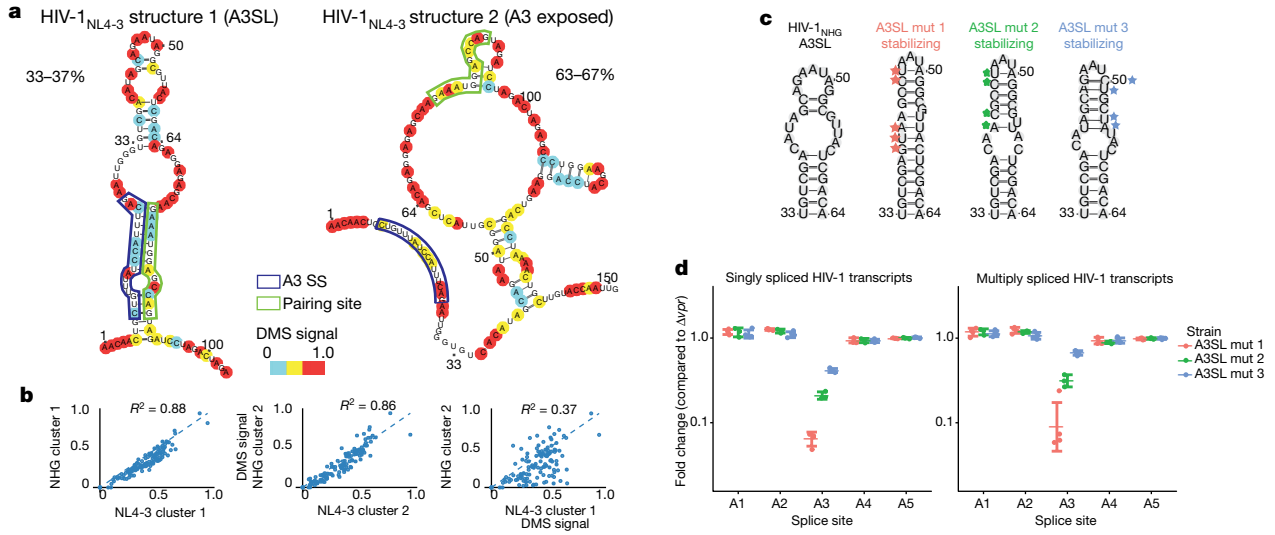


Fig. 3 | Alternative RNA structures at the A3 splice acceptor site regulate splice site use. a, Structural models of the A3ss from CD4$^+$ T cells infected with HIV-1_{NL4-3}, made from the clustering outputs of DREEM. Proportions of each cluster are a range from $n = 4$ experiments (1 HIV-1_{NL4-3} and 3 HIV-1_{NHG}). Nucleotides are colour-coded by normalized DMS signal. The splice site is highlighted in a blue box; a region that base-pairs to the splice site is shown in green. **b**, Scatter plots comparing alternative structures between CD4$^+$ T cells infected with HIV-1_{NL4-3} and HEK293T cells transfected with HIV-1_{NHG}, from $n = 2$ experiments. The blue dotted line is the identity line; R^2 is Pearson's R^2. **c**, Mutant design. All mutants were predicted to thermodynamically stabilize the A3SL. **d**, Fold change in splice site use, compared to Δ*vpr* HIV-1_{NHG} ($n = 4$ for A3SL mut 1 and mut 3; $n = 3$ for A3SL mut 2). Left, splice site use of singly spliced transcripts. Right, multiply spliced transcripts. Data are reported as points, with mean and error bars representing the s.d.

effect on splicing—increased A3ss use relative to the parental strain (A3SL mut 4) (Extended Data Fig. 7a, b). To further test the inhibitory role of A3SL, we designed a compensatory mutant to shift the population towards the A3SL in the sequence context of the A3SL mut 4. Consistent with A3SL inhibiting splicing, the compensatory mutant (A3SL mut 5) uses A3ss about 10-fold less frequently than Δ*vpr* HIV-1_{NHG} (Extended Data Fig. 7b). The percentage of the A3SL cluster for each mutant had an inverse relationship with the overall use of the A3 splice acceptor site (Extended Data Fig. 7c). To understand the origin of the increase in splicing, we probed A3SL mut 1 and A3SL mut 4 and found that these mutations resulted in the formation of an unanticipated alternative structure in cluster 2 of both mutants (Extended Data Fig. 8a, b). Cluster 2 was present at 35% for A3SL mut 1 and 53% for A3SL mut 4. This result demonstrates that thermodynamic predictions alone are incomplete. The unanticipated structures alter the accessibility of multiple nearby protein-binding sites. These results indicate that the intrinsic ability of RNA to form alternative structures can regulate splicing either by directly occluding U2AF binding sites or by modifying the accessibility of nearby splicing enhancer and silencer elements; the net effect of this regulation results in up to about a 100-fold change in abundance of HIV-1 Tat transcripts.

To test whether the formation of alternative structures is a general property of HIV-1 RNA, we prepared a genome-wide DMS-MaPseq dataset from HEK293T cells transfected with HIV-1_{NHG} (Extended Data Fig. 9a). We used DREEM clustering on overlapping windows spanning the entire genome and applied a stringent Bayesian information criteria (BIC) test to determine whether the data could be separated into two distinct structure signals[24]. Notably, both the RRE and A3ss match the results obtained by specific PCR with reverse transcription (RT–PCR) (Extended Data Fig. 9b, c).

More than 90% of windows with coverage of >100,000 sequencing reads passed the BIC test for 2 clusters, indicating the presence of heterogeneity in RNA structure across the entire HIV-1 genome. We quantified the variability in reactivity of residues in each window using the Gini index metric, which is used to estimate the stability of the RNA structure[25]. A Gini index close to zero indicates a relatively even distribution of DMS modifications, and occurs when RNA is unfolded or when RNA structure is highly heterogeneous. A Gini index close to one occurs when a subset of residues is strongly protected from DMS, and indicates a highly stable structure. We also computed a Pearson's correlation coefficient for all windows that had alternative structures to measure how different the two structures were from each other. The low Pearson correlation ($R^2 < 0.3$) and low Gini index (<0.5) indicate that that relatively unstable, alternative structures form across the entire genome (Fig. 4a)—including alternative conformations for a conserved structure[26] in the 4-kilobase (kb) *gag-pro-pol* region (Extended Data Fig. 9d), which is present exclusively in unspliced transcripts. The smallest minor cluster that we observed was present at 20% and was located in the *env* coding region (Extended Data Fig. 10a).

The widespread alternative structure of the HIV-1 genome stood in contrast to the small nuclear RNA U1 probed in vivo and U4 and U6 core-domain RNA probed in vitro, both of which exhibited minimal heterogeneity (Extended Data Fig. 10b, c). These RNAs have stable structures, as previously determined by X-ray crystallography[27] and nuclear magnetic resonance[28], respectively. As a control against over-clustering, we simulated reads on the basis of the HIV-1 population-average DMS signal with no relationship between mutations, and observed no regions that passed the BIC test for two clusters (Extended Data Fig. 10d). As expected, we observed an inverse relationship between the Gini index and Shannon entropy, an alternative measure of RNA structure (Extended Data Fig. 11a, b). We used the whole-genome data to identify previously validated structures such as the transcription activation region, which was detected in one conformation (Extended Data Fig. 11c). We found structure heterogeneity at most splice sites, including A4a, A4b, A4c and A5 (Extended Data Fig. 11d). Together, these results suggest splice-site occlusion as a general mechanism through which HIV-1 tunes alternative splicing.

In summary, our results indicate that the thermodynamic ability of RNA to form alternative conformations at critical splice sites enables HIV-1 to express different genes from the same primary transcript. This may be necessary from an evolutionary perspective for HIV-1 to set an upper limit for splice-site use independent of splice enhancer and suppressor recognition. Splicing repression by RNA structure could ensure that a fraction of molecules remain unspliced, which is essential for packaging and transmitting the full-length HIV-1 genome. Finally, DREEM clustering permits the study of alternative RNA structures at

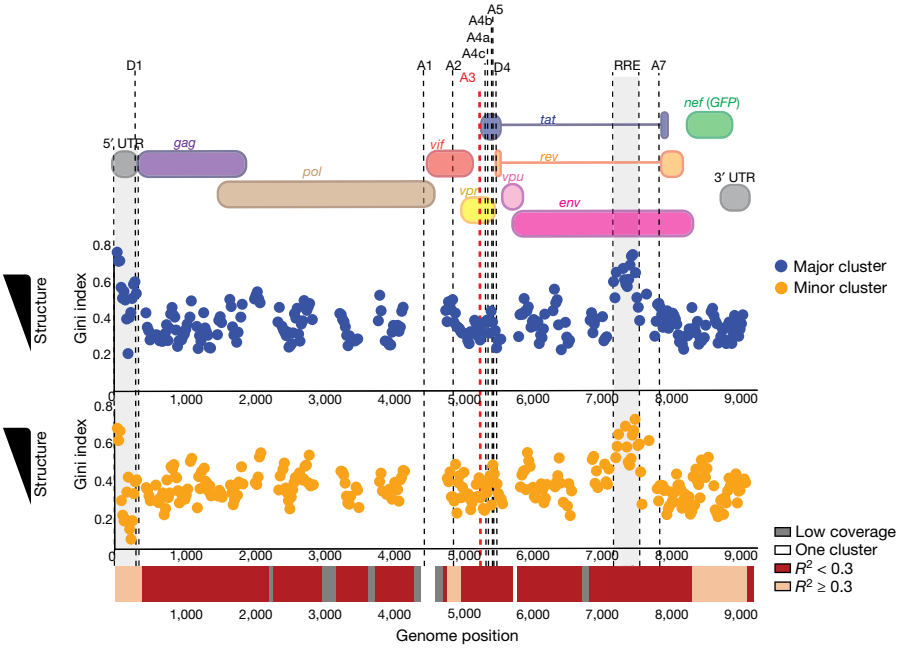


Fig. 4 | Landscape of heterogeneity in HIV-1 RNA. HIV-1 genome organization, highlighting the UTR, coding regions, major splice donor and acceptor sites and RRE overlaid on a structural variability plot for the library generated from HEK293T cells transfected with HIV-1_{NHG}. Each dot represents an 80-nt window of DMS-MaPseq data used for DREEM with a maximum of 2 clusters from $n = 1$ experiment. The cluster for each window with a higher Gini coefficient is plotted on top in orange, and the cluster with a lower Gini coefficient is plotted on bottom in blue. A heat map comparing the Pearson's R^2 for the two clusters is below the Gini coefficient. Windows without sufficient coverage for clustering (<100,000 reads) are in grey. Windows that did not pass the BIC test for more than one cluster are in white. A Pearson's R^2 value measures the similarity in the DMS signal between each pair of clusters identified by DREEM. The most divergent clusters, with $R^2 < 0.3$, are in red; clusters with $R^2 \geq 0.3$ are in orange.

single-nucleotide resolution in living cells. The DREEM approach has wide range of potential applications, including elucidating the role of RNA structure in human alternative splicing—where changes of splice site use of as little as twofold are associated with multiple diseases[29,30].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2253-5.

1. 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).
2. Ocwieja, K. E. et al. Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic Acids Res.* **40**, 10345–10355 (2012).
3. Takata, M. A. et al. Global synonymous mutagenesis identifies *cis*-acting RNA elements that regulate HIV-1 splicing and replication. *PLoS Pathog.* **14**, e1006824 (2018).
4. Watts, J. M. et al. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711–716 (2009).
5. Sherpa, C., Rausch, J. W., Le Grice, S. F., Hammarskjold, M. L. & Rekosh, D. The HIV-1 Rev response element (RRE) adopts alternative conformations that promote different rates of virus replication. *Nucleic Acids Res.* **43**, 4676–4686 (2015).
6. Warf, M. B. & Berglund, J. A. Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem. Sci.* **35**, 169–178 (2010).
7. Shepard, P. J. & Hertel, K. J. Conserved RNA secondary structures promote alternative splicing. *RNA* **14**, 1463–1469 (2008).
8. Singh, N. N., Lee, B. M. & Singh, R. N. Splicing regulation in spinal muscular atrophy by an RNA structure formed by long-distance interactions. *Ann. NY Acad. Sci.* **1341**, 176–187 (2015).
9. Huthoff, H. & Berkhout, B. Two alternating structures of the HIV-1 leader RNA. *RNA* **7**, 143–157 (2001).
10. Abbink, T. E., Ooms, M., Haasnoot, P. C. & Berkhout, B. The HIV-1 leader RNA conformational switch regulates RNA dimerization but does not regulate mRNA translation. *Biochemistry* **44**, 9058–9066 (2005).
11. Zubradt, M. et al. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Methods* **14**, 75–82 (2017).
12. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
13. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
14. Spasic, A., Assmann, S. M., Bevilacqua, P. C. & Mathews, D. H. Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Res.* **46**, 314–323 (2018).
15. Homan, P. J. et al. Single-molecule correlated chemical probing of RNA. *Proc. Natl Acad. Sci. USA* **111**, 13858–13863 (2014).
16. Sengupta, A., Rice, G. M. & Weeks, K. M. Single-molecule correlated chemical probing reveals large-scale structural communication in the ribosome and the mechanism of the antibiotic spectinomycin in living cells. *PLoS Biol.* **17**, e3000393 (2019).
17. Ding, Y. & Lawrence, C. E. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.* **31**, 7280–7301 (2003).
18. Halvorsen, M., Martin, J. S., Broadaway, S. & Laederach, A. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.* **6**, e1001074 (2010).
19. Wan, Y. et al. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–709 (2014).
20. Tian, S., Kladwang, W. & Das, R. Allosteric mechanism of the *V. vulnificus* adenine riboswitch resolved by four-dimensional chemical mapping. *eLife* **7**, e29602 (2018).
21. Lemay, J. F. et al. Comparative study between transcriptionally- and translationally-acting adenine riboswitches reveals key differences in riboswitch regulatory mechanisms. *PLoS Genet.* **7**, e1001278 (2011).
22. Zaug, A. J. & Cech, T. R. Analysis of the structure of *Tetrahymena* nuclear RNAs in vivo: telomerase RNA, the self-splicing rRNA intron, and U2 snRNA. *RNA* **1**, 363–374 (1995).
23. Emery, A., Zhou, S., Pollom, E. & Swanstrom, R. Characterizing HIV-1 splicing by using next-generation sequencing. *J. Virol.* **91**, e02515-16 (2017).
24. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
25. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*. *Nature* **505**, 701–705 (2014).
26. Liu, Y. et al. The roles of five conserved lentiviral RNA structures in HIV-1 replication. *Virology* **514**, 1–8 (2018).
27. Kondo, Y., Oubridge, C., van Roon, A. M. & Nagai, K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5′ splice site recognition. *eLife* **4**, e04986 (2015).
28. Cornilescu, G. et al. Structural analysis of multi-helical RNAs by NMR-SAXS/WAXS: application to the U4/U6 di-snRNA. *J. Mol. Biol.* **428** (5 Pt A), 777–789 (2016).
29. Xiong, H. Y. et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
30. Faustino, N. A. & Cooper, T. A. Pre-mRNA splicing and human disease. *Genes Dev.* **17**, 419–437 (2003).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

DREEM clustering description

Definitions of symbols: N, total number of reads; D, length of region of interest in the reference; $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, set of all observed reads; S, set of all allowed (observable) reads; K, number of clusters; π_k, mixing proportion of cluster k. $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ such that $\sum_{k=1}^{K} \pi_k = 1$. $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$ in which $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{ki}, \ldots, \mu_{kD})\ \forall\, i = 1, \ldots, D$ is the mutation profile of cluster k and in which μ_{ki} is the mutation rate of base i in cluster k. y_{nk}, the latent Boolean variable representing the assignment of read n to cluster k. z_{nk}, the expectation of y_{nk}, or the probability that read n belongs to cluster k. i, α, nucleotide index.

The sequencing data from a sample were mapped to the corresponding reference genome using the Bowtie2 aligner[31]. The data observed X consists of N reads $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, each containing D nucleotides. Each read $\boldsymbol{x}_n \in \boldsymbol{X}$ represents a distinct RNA molecule that was DMS-modified, reverse-transcribed and amplified. The DMS modifications are read out as mutations. A read $\mathbf{x}_n$ can then be represented as a vector of D bits $(x_{n1}, \ldots, x_{nD})$ or a 'bit vector' in which $x_{ni} = 1$ if base x_{ni} is mutated, or 0 otherwise.

As DMS modification is far from saturating (that is, not every accessible base of a single molecule is modified), each open base in an RNA molecule has only a small probability (2–10%, depending on the DMS concentrations used) of being modified. External factors unrelated to the secondary structure (such as 3D conformation or local chemical environment) will affect this probability. As a consequence of this, a distinct mutation probability μ will be associated with each base of the read. We assume the mutation probabilities are independent from each other. This assumption allows us to consider each read as a random draw from a Bernoulli mixture model. In the event that the RNA molecules assume more than one structure, each structure will appear in the data as a collection of reads (that is, a cluster), characterized by its own Bernoulli mixture model.

If K is the number of structures present in our sample, then the model is parameterized by: the mutation probabilities $\mu = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$, in which $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kD})$ are the mutation probabilities of cluster k, and the mixing proportions $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ of the K clusters, in which π_k quantifies the proportion of reads that belong to cluster k.

The expectation–maximization algorithm used by DREEM for clustering assumes a Bernoulli mixture model[12]. Therefore, the probability (Pr) of a base not being mutated in cluster k is: $\Pr(x_{ni} = 0|\boldsymbol{\mu}_k) = 1 - \mu_{ki}$, and the probability of a base being mutated in cluster k is: $\Pr(x_{ni} = 1|\boldsymbol{\mu}_k) = \mu_{ki}$. Therefore, the Bernoulli mixture model gives us the probability of observing a read $\mathbf{x}_n$ from cluster k as:

$$\Pr(\mathbf{x}_n|\boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_{ni}}(1-\mu_{ki})^{1-x_{ni}} \quad (1)$$

We observed that in DMS-MaPseq data, reads that contain mutations within three bases of each other are very rare, and occur at a frequency close to the sequencing error rate (Extended Data Fig. 2); that is, the bit vectors 001001000, 001010000 and 001100000 are greatly underrepresented. This is probably due to the reverse transcriptase falling off the template when encountering adjacent methylations. Truncated reads do not get amplified during PCR, and therefore are not represented when sequenced. To account for this bias, we remove all rare reads containing mutations within three bases of each other and we compute $\boldsymbol{S}$, the set of all reads with allowable mutations in $\{0,1\}^D$ that can be sequenced. Therefore, equation (1) is modified as follows:

$$\Pr(\mathbf{x}_n|\boldsymbol{\mu}_k) = \frac{\prod_{i=1}^{D}(\mu_{ki})^{x_{ni}}(1-\mu_{ki})^{1-x_{ni}}}{\sum_{x' \in S}\prod_{i=1}^{D}(\mu_{ki})^{x'_i}(1-\mu_{ki})^{1-x'_i}}.$$

In the initial step of the expectation–maximization algorithm, the model parameters $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$ are randomly initialized. After the initialization of the parameters, the expectation step and the maximization step are executed one after the other in a loop until the log likelihood converges.

Two calculations are made in the expectation step: first, the responsibilities of the cluster are computed—that is, the reads are assigned probabilistically to clusters:

$$z_{nk} = \frac{\Pr(\mathbf{x}_n|\boldsymbol{\mu}_k)\pi_k}{\sum_{j=1}^{K}\Pr(\mathbf{x}_n|\boldsymbol{\mu}_j)\pi_j}.$$

Here z_{nk} is the probability that read n belongs to cluster k. It can also be defined as the posterior probability, or responsibility, of cluster k given read n. Second, the expected complete-data log-likelihood of observing the data X and latent variables $Y = \{Y_{nk}\}$ given the model parameters is computed:

$$\mathbb{E}_{Y\sim Z} \ln \Pr(X, Y|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} x \sum_{k=1}^{K} z_{nk} \ln\{\pi_k \Pr(\mathbf{x}_n|\boldsymbol{\mu}_k)\}$$

In the maximization step, the model parameters are re-estimated by maximizing the expected value of the likelihood with respect to the parameters $\{\pi_k\}$ and $\{\mu_{ki}\}$. The mixing proportion of each cluster is then updated, using:

$$\pi_k = \frac{\sum_{n=1}^{N} z_{nk}}{N}$$

The mutation profile μ_k of each cluster is then updated by solving the following system of equations for each k:

$$\frac{\sum_{x\in S}\mathbf{x}_\alpha \prod_{i=1}^{D}(\mu_{ki})^{\mathbf{x}_i}(1-\mu_{ki})^{1-\mathbf{x}_i}}{\sum_{x\in S}\prod_{i=1}^{D}(\mu_{ki})^{\mathbf{x}_i}(1-\mu_{ki})^{1-\mathbf{x}_i}} = \frac{\sum_{n=1}^{N} z_{nk}x_{n\alpha}}{\sum_{n=1}^{N} z_{nk}} \quad \forall\alpha$$

These equations are derived by setting the derivatives of the expected complete-data log-likelihood function to zero.

After the expectation–maximization clustering algorithm has finished running, the reactivity of the bases in each cluster is given as input to RNAstructure[13] for secondary structure prediction. The DMS signal is normalized such that the median of the top ten most-reactive positions is set to 1.0. To protect from spurious outliers, we use 90% winsorization, effectively capping the reactivity at 1.0. Final visualizations of RNA secondary structure were created with VARNA[32].

The parameters used by the DREEM pipeline were as follows: the minimum number of iterations of the expectation–maximization algorithm to run before checking for convergence of the likelihood (num_its), 300; the number of expectation–maximization algorithm runs (num_runs), 10. num_runs independent runs of the expectation–maximization algorithm are carried out to ensure that the results from the algorithm are robust to the initialization of the model parameters and are repeatable.

The convergence threshold (conv_thresh) is 1. The expectation–maximization algorithm is stopped when $\log(\text{likelihood})_{\text{iteration}=n+1} - \log(\text{likelihood})_{\text{iteration}=n} < \text{conv\_thresh}$ after num_its iterations have been completed.

The signal threshold (sig_thresh) is 0.005. Only mutation rates greater than sig_thresh are considered. All bases with a population-average mutation rate less than sig_thresh are set to 0 in every bit vector.

We used $\text{BIC} = \log(N) \times D \times K - 2\log(\text{likelihood})$. To test for over fitting the data, we checked whether the expectation–maximization algorithm

passes two clusters by using the BIC test. If $BIC_{K=2} > BIC_{K=1}$, the algorithm stops. Otherwise, the algorithm moves on to $K = 3$.

Bit vectors are filtered out if they do not satisfy one of the following four criteria: informative bits threshold (info_thresh) of 0.05–0.2. We set x_{ni} to '.' if base i is not covered by read x_n and to '?' if it is of low quality (defined as having a Phred quality score of less than 20). If the fraction of non-informative bits ('.', '?' and *N*) in the bit vector is greater than info_thresh, the bit vector is removed. After this filtering, all the non-informative bits are set to 0 in the remaining bit vectors. We set a maximum number of mutations such that if the number of mutations in the bit vector is greater than three times the standard deviation of the mutation distribution per read, the bit vector is removed. Invalid bit vectors represent rare occurrences of bit vectors with adjacent mutations (within 3 nt) are considered to be part of background noise (Extended Data Fig. 2) and are filtered out. There are also rare instances in which a bit vector consisted of a mutation (1) right next to a non-informative base such as '.' or '?'. These reads were also filtered out. Because DMS modifies only A and C, these constitute informative bases: mutations at T and G are set to 0.

Cell lines

HEK293T cells were obtained from ATCC. The cells tested negative for mycoplasma by LookOut Mycoplasma PCR Detection kit (Millipore-Sigma). The cells were maintained in Dulbecco's Modified Eagle Medium (ThermoFisher Scientific) supplemented with 10% heat-inactivated fetal bovine serum (ThermoFisher Scientific) and 100 U/ml penicillin–streptomycin (ThermoFisher Scientific).

Plasmid construction

HIV-1_{NL4-3} infectious molecular clone (pNL4-3) was obtained from the NIH AIDS reagent programme[33]. HIV-1_{NHG} is a full-length HIV-1 proviral plasmid, modified to replace a non-essential gene (*nef*) with *GFP* (GenBank accession code: JQ585717.1). A Vpr-truncated derivative (Δ*vpr* HIV-1_{NHG}) was constructed by generating an overlapping PCR with a C-to-T mutation, and thus a stop codon after Vpr amino acid 20. This PCR product inserted into HIV-1_{NHG} using AgeI and SalI. All of the A3 splice site mutants were generated via overlapping PCR and inserted into a Δ*vpr* HIV-1_{NHG}.

$CD4^+$ T cell isolation

Apheresis leukoreduction collars, obtained from the Brigham and Women's Hospital Crimson Core, were used to isolate peripheral blood mononuclear cells by lymphocyte separation medium (ThermoFisher Scientific) density centrifugation. $CD4^+$ T cells were isolated by negative selection using EasySep Human $CD4^+$ T cell Enrichment Kit (StemCell Technologies). $CD4^+$ T lymphocytes were cultured at a density of approximately 1 million cells per millilitre in RPMI-1640 (ThermoFisher Scientific) supplemented with 10% fetal bovine serum and 100 U/ml penicillin–streptomycin.

DMS modification of in vitro-transcribed RNA

gBlocks were obtained from IDT for the HIV-1 RRE, RRE mut A and mut B, control structure 1, control structure 2 and adenoriboswitch. HIV-1 RRE and its mutants correspond to nucleotides 7,759–7,990 based on HIV-1 vector pNL4-3 (GenBank accession code: AF324493.1). Adenosine deaminase (*add*) riboswitch corresponds to nucleotides 1,590,535–1,590,663 of *Vibrio vulnificus* strain (GenBank accession code: CP037932.1). The U4 and U6 core-domain RNA construct is based on the interface of the U4 and U6 snRNA (GenBank accession code: 2N7M_X). The gblock also contain 20-nt T7 RNA polymerase promoter sequence (TTCTAATACGACTCACTATA) on the 5′ end and a 23-nt sequence (CCGGAGTCGAGTAGACTCCAACA) on the 3′ end. The region of interest was amplified by PCR with a forward primer that contained the T7 promoter sequence. The PCR product was used for T7 Megascript in vitro transcription (ThermoFisher Scientific) according to manufacturer's instructions. Subsequently, 1 µl Turbo DNase I (ThermoFisher Scientific) was added to the reaction and incubated at 37 °C for 15 min. The RNA was purified using RNA Clean and Concentrator -5 kit (Zymo). Between 1 and 2 µg of RNA was denatured at 95 °C for 1 min. On the basis of the DMS concentration used in the next step, 300 mM sodium cacodylate buffer (Electron Microscopy Sciences) with 6 mM $MgCl^{2+}$ was added so that the final volume was 100 µl. The RNA was refolded by incubating for 20 min at 37 °C. Then, 0.25–2.5% DMS (Millipore-Sigma) was added and incubated at 37 °C for 5 min while shaking at 500 r.p.m. on a thermomixer. The DMS was neutralized by adding 60 µl β-mercaptoethanol (Millipore-Sigma). The RNA was purified using RNA Clean and Concentrator -5 kit. For in vitro transcription of *add* riboswitch samples, one set of samples were incubated with 5 mM adenine during the refolding stage at 37 °C.

$CD4^+$ T cell infection and DMS modification

Fifteen million $CD4^+$ T cells were activated by treatment with culture medium containing 10 µg/ml PHA (Millipore-Sigma) and 100 U/ml IL-2 (ref. [34]) (NIH AIDS reagent programme; discontinued) for 72 h. The cells were pelleted and infected in a small volume with supernatant from HEK293T cells transfected with pNL4-3 for 2 h, then culture medium is added to achieve a concentration of about 1 million cells per ml. Subsequently, 72 h after infection, the supernatant was filtered with a 0.22-µm filter (Millipore-Sigma) and centrifuged at 28,000*g* for 1 h at 4 °C to pellet virions. The cells were washed and resuspended in 15 ml of medium and placed on a thermomixer at 37 °C. Then, 200 µl DMS, or about 1.3% v/v, (Millipore-Sigma) was added and the cells were incubated for 10 min while shaking at 800 r.p.m. DMS was neutralized by adding 30 ml PBS (ThermoFisher Scientific) with 30% β-mercaptoethanol. The cells were centrifuged at 1,000*g* for 5 min at 4 °C. The cells were washed twice by resuspending the pellet with 15 ml PBS with 30% β-mercaptoethanol and centrifugation to pellet. After washes, the pellet was resuspended in 1 ml Trizol (ThermoFisher Scientific) and RNA was extracted following the manufacturer's specifications. The virions were resuspended in 400 µl PBS with 10 mM Tris pH 7 and 3 mM $MgCl^{2+}$. Next, 40 µl DMS was added and the virions were incubated at 37 °C on a thermomixer while shaking at 800 r.p.m. for 10 min. The DMS was neutralized with 400 µl β-mercaptoethanol and the RNA was purified using RNA Clean and Concentrator -5 kit. For unmodified RNA, 15 million $CD4^+$ T cells were isolated and infected in a small volume with supernatant from HEK293T cells transfected with pNL4-3 for 2 h, then culture medium is added to achieve a concentration of about 1 million cells per ml. Next, 72 h after infection, the supernatant was filtered with a 0.22-µm filter and virions were pelleted from the supernatant by centrifugation at 28,000*g* for 1 h at 4 °C and resuspended in 1 ml Trizol. The cells were pelleted and resuspended in 1 ml Trizol. RNA was extracted following the manufacturer's instructions.

HEK293T cell transfection and DMS modification

Nine hundred thousand cells per well were seeded on a six-well plate and incubated overnight. Next, 2 µg of plasmid DNA (NL4-3, NHG or mutant) per well was transfected into the cells using X-tremeGENE 9 (Millipore-Sigma) following the manufacturer's instructions, and incubated for 48 h. After incubation, virions were collected from the supernatant and DMS-modified as in '$CD4^+$ T cell Infection and DMS modification'. The cells were washed with PBS, and 2 ml medium with about 1.3% v/v DMS was added to each well. The plates were incubated at 37 °C for 4 min. The medium containing DMS was immediately removed and replaced with PBS with 30% β-mercaptoethanol. Cells were scraped and centrifuged at 1,000*g* for 5 min at 4 °C. The pellet was resuspended in PBS and centrifuged to pellet twice. The pellet was resuspended in 1 ml Trizol and RNA was extracted following the manufacturer's specifications. For unmodified RNA, 900,000 HEK293T cells were seeded on a 6-well plate and transfected 2 µg of plasmid DNA (NL4-3, NHG or mutant) per well into the cells using X-tremeGENE

9 following the manufacturer's instructions. Then, 48 h after transfection, the supernatant was filtered with a 0.22-μm filter and virions were pelleted from the supernatant by centrifugation at 28,000*g* for 1 h at 4 °C and resuspended in 1 ml Trizol. The cells were trypsinized, washed and resuspended in 1 ml Trizol. RNA was extracted following the manufacturer's instructions.

RT–PCR with DMS-modified RNA from cells or in vitro transcription

Between 1 and 3 μg of RNA per reaction was used as the input for rRNA subtraction. First, 1 μl rRNA subtraction mix (3 μg/μl) and 2.5 μl 5× hybridization buffer (1 M NaCl, 500 mM Tris-HCl pH 7.5) were added to each reaction, and final volume was then adjusted with water to 12.5 μl. The samples were incubated at 68 °C and the temperature was reduced by 1 °C/min until the reaction was at 45 °C. Next, 5 μl RNase H buffer and 2 μl hybridase thermostable RNase H (Lucigen) were added and water was added until the final volume was 40 μl. The samples were incubated at 45 °C for 30 min. The RNA was cleaned with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of fragments >200 nt and eluted in 45 μl water. Then, 5 μl Turbo DNase buffer and 1 μl Turbo DNase (ThermoFisher Scientific) were added to each reaction and incubated for 30 min at 37 °C. Then, 5.1 μl DNase inactivation reagent (ThermoFisher Scientific) was added and incubated 5 min at room temperature with intermittent manual mixing. The RNA was cleaned with RNA Clean and Concentrator -5 following instructions for recovery of fragments >200 nt and eluted in 15 μl water. For reverse transcription, 1 μl of RNA was added to 3.5 μl water, 2 μl 5× first strand buffer (ThermoFisher Scientific), 1 μl 10μM reverse primer, 1 μl dNTP, 0.5 μl 0.1M DTT, 0.5 μl RNaseOUT and 0.5 μl TGIRT-III (Ingex). The reverse-transcription reaction was incubated at 57 °C for 1.5 h, followed by 5 min at 80 °C. To degrade the RNA, 1 μl RNase H (New England Biolabs) was added to the reverse-transcription reaction and incubated for 20 min at 37 °C. PCR was performed to amplify the samples using either Advantage HF 2 DNA polymerase (Takara) or Phusion (NEB) for 25–30 cycles according to the manufacturer's specifications. The PCR product was purified by QIAquick PCR purification (Qiagen) and sequenced either on MISeq or iSeq100 (Illumina) to produce either 100-nt single-end reads or 150×150-nt paired-end reads.

Library generation with DMS-modified RNA for HIV-1 genome RNA structure

Extracted DMS-modified RNA from HEK293T cells transfected with the NHG plasmid (10 μg) was split into three reactions for the first step of RNase H-based rRNA subtraction. The steps for RNase H and DNase treatment mentioned in 'RT–PCR with DMS-modified RNA from cells or in vitro transcription' were followed. After DNase treatment, the 3 reactions were eluted in 8.5 μl water and combined. An additional rRNA subtraction step was performed using the RiboZero Human/Mouse/Rat rRNA removal kit (Illumina; discontinued) according to the manufacturer's specifications. After RiboZero, the RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of fragments >200 nt and eluted in 10 μl water. The RNA was fragmented using the RNA Fragmentation kit (ThermoFisher Scientific) with a fragmentation step of 45 s at 70 °C. The RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of all fragments and eluted in 6.5 μl water. Then, 1 μl CutSmart buffer (New England Biolabs), 1.5 μl shrimp alkaline phosphatase (New England Biolabs) and 1 μl RNaseOUT (ThermoFisher Scientific) were added and incubated at 37 °C for 1 h to dephosphorylate the RNA. Subsequently, 6 μl 50% PEG-800 (New England Biolabs), 2.2 μl 10× T4 RNA ligase buffer (New England Biolabs), 2 μl T4 RNA ligase, truncated KQ (England Biolabs) and 1 μl linker were added to the reaction and incubated for 18 h at 22 °C. The RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of all fragments and eluted in 15 μl water. Excess linker was degraded by adding 2 μl 10× RecJ buffer (Lucigen), 1 μl RecJ exonuclease (Lucigen), 1 μl 5′ deadenylase (New England Biolabs) and 1 μl RNaseOUT, then incubating for 1 h at 30 °C. The RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of fragments >200 nt and eluted in 11 μl water. For reverse transcription, 1 μl reverse-transcription primer, 1 μl 0.1M DTT, 4 μl 5× first strand buffer, 1 μl dNTP, 1 μl RNaseOUT and 1 μl T-GIRT III were added and the sample was incubated for 2 h at 65 °C. RNA was degraded by adding 1 μl 4 N NaOH and incubating at 95 °C for 3 min. The reverse-transcription product was mixed with an equal volume 2× Novex TBE-urea sample buffer (ThermoFisher Scientific) and run on a 10% TBE-urea gel (ThermoFisher Scientific) and the approximately 300–400-nt product was extracted. The purified single-strand DNA was circularized using the CircLigase ssDNA ligase kit (Lucigen). Then, 2 μl of the circularized product was used for PCR using Phusion. The sample was run for a maximum of 14 cycles. Following PCR, the product was run on an 8% TBE gel and the approximately 350–450-nt product was gel-extracted. The final PCR product was quantified by Bioanalyzer (Agilent). The product was then sequenced by Novaseq S4 (Illumina) to produce 150×150-nt paired-end reads. The same library generation protocol was followed for in vitro-transcribed and DMS-modified U4 and U6 core-domain with some modifications. The starting amount was 250 ng of RNA as part of a pool of RNA totalling 4 μg. No fragmentation and no rRNA removal were performed.

HIV-1 splice junction use analysis

Splice analysis was performed according to a previously written protocol[24]. In brief, two separate RT–PCR reactions were performed with 2 μg total unmodified RNA from HEK293T cells transfected with plasmid containing HIV-1_{NHG}, Δ*vpr* HIV-1_{NHG} or HIV-1 mutants. One reaction was designed to reverse-transcribe all HIV-1 multiply spliced products with a reverse primer that spans the D4A7 splice junction. The second reaction is designed to reverse-transcribe HIV-1 singly spliced mRNA with a reverse primer that lies in the *env* intron. The forward primer used in both PCR reactions is located upstream of D1. Reverse transcription was performed with SuperScript III (Thermo Fisher Scientific) at 55 °C for 1 h followed by 15 min at 70 °C. RNA was degraded by adding 1 μl RNase H and incubating at 37 °C for 20 min. The cDNAs were then purified with Agencourt RNACleanX beads at a ratio of 2:1 (Beckman Coulter). Two successive rounds of PCR were used to add adapters for sequencing using the KAPA robust PCR kit (KAPA Biosystems). The first PCR uses with a forward primer that is located in the shared upstream D1 sequence that also has an adaptor. The second round adds the universal adaptor and Illumina- ndexed sequencing primers. The PCR products were then sequenced by Illumina Miseq, 300× 300-nt paired-end reads.

Statistical methods

Statistical analysis of DREEM clusters was quantified by Pearson's correlation. R^2 and *P* values of Pearson's correlation are reported.

Library linker and primers

All oligonucleotides were ordered from IDT. The stem A and stem C T7 forward primer: TAATACGACTCACTATAGAAAGGATCGG; stem A and stem C T7 reverse primer: ATCCCAGCGCGTGGTGCA; stem A and stem C reverse transcription primer: ATCCCAGCGCGTGGTGCA; stem A and stem C PCR forward primer: GAAAGGATCGGAAGACTCCACAG; stem A and stem C PCR reverse primer: ATCCCAGCGCGTGGTGCA; *add* riboswitch T7 forward primer: TTCTAATACGACTCACTATAGGAC ACGACTCGAGTAGAGTCG; *add* riboswitch forward primer: GAC ACGACTCGAGTAGAGTCG; *add* riboswitch reverse primer: TGTTGGA GTCTACTCGACTCCGGT; HIV-1 RRE T7 forward primer: TAATACGAC TCACTATAGGAGCTTTGTTCC; HIV-1 RRE T7 reverse primer: GGAGCTGT TGATCCTTTAGGTATCTTTC; HIV-1 RRE RT primer: GGAGCTGT TGATCCTTTAGGTATCTTTC; HIV-1 RRE PCR forward primer: GGAG CTTTGTTCCTTGGGTTCTTGG; HIV-1 RRE PCR reverse primer: GGA

GCTGTTGATCCTTTAGGTATCTTTC; HIV-1 A3 PCR forward primer: TGAAACTTACGGGGATACTTGGGCAGGA; HIV-1_{NL4-3} A3 PCR and reverse transcription reverse primer: GAAGCTTGATGAGTCTGACTGTTCTGA TGAGC; HIV-1_{NHG} A3 PCR and reverse transcription reverse primer: CTTCGTCGCTGTCTCCGCTTCTTCC.

To generate Δ*vpr* HIV-1_{NHG}, we used: HIV-1_{NL4-3} *ageF*: AGCTAGAACTGG CAGAAAACAGGGAGATTC; NL *SalIR*: CCATTTCTTGCTCTCCTCTGT CGAGTAACGC; Δ*vprS*: GGAAACTGACAGAGGACAGATGGAATAAGCCCC AGAAGACC; and Δ*vprAS*: GGTCTTCTGGGGCTTATTCCATCTGTCCTCTGT CAGTTTCC.

To generate A3 splice site mutants, we used: NL 5599F: CATACAA TGAATGGACACTAGAGCTTTTAG; NL *BamHIR*: GTCCCAGATAAGTG CCAAGGATCCGTT; A3SL mut 1 S: TCCATTTCAGAATTGGGTGTCGAGTA AGCCTAATAGGCGTTACTCGACAGAGGA; A3SL mut 1 AS: TCCTCTGTCG AGTAACGCCTATTAGGCTTACTCGACACCCAATTCTGAAATGGA; A3SL mut 2 S: GAATTGGGTGTCGACAACGCCTAATAGGCGTTACTCG AC; A3SL mut 2 AS: GTCGAGTAACGCCTATTAGGCGTTGTCGACACC CAATTC; A3SL mut3 S: GGTGTCGACATAGCAGAATCTGCTATACTCGA CAGAGGAGAGCAA; A3SL mut 3 AS: GGTGTCGACATAGCAGAATCTGC TATACTCGACAGAGGAGAGCAA; A3SL mut 4 S: TCAGAATTGGGTGTCGAA ACAGCGAAATAGGCGTTACTCGACAGA; A3SL mut 4 AS: TCTGTCGAG TAACGCCTATTTCGCTGTTTCGACACCCAATTCTGA; A3SL mut 5 S: TCA GAATTGGGTGTCGAAACAGCGAAATTCGCGTGTTTCGACAGAGGAGAG CAA; A3SL mut 5 AS: TTGCTCTCCTCTGTCGAAACACGCGAATTTCGCTGT TTCGACACCCAATTCTGA.

The library generation linker was: /5rApp/TCNNNNNNNNNN NAGATCGGAAGAGCGTCGTGTAGGGAAAGA/3ddC/. The library generation reverse-transcription primer was: /5Phos/AGATCGGA AGAGCACACGTCTGAACTCCAG/iSp18/TCTTTCCCTACACGACGCTC TTCCGATCT. The library generation forward PCR primer was: CAAGC AGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGC TC (in which X denotes any base). The library generation reverse PCR primer was: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA CGACGCTC.

For splice analysis, the following primers were used. Multiply spliced reverse primer: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTN NN NNNNNNNNNNNCAGTTCGGGATTGGGAGGTGGGTTGC; singly spliced reverse primer: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT NN NNNNNNNNNNNNGTACTATAGGTTGCATTACATGTACTACTTAC; PCR round 1 forward primer: GCCTCCCTCGCGCCATCAGAGATGTGT ATAAGAGACAGNNNNTGCTGAAGCGCGCACGGCAAG; PCR round 2 reverse primer: CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGG AGTTCAGACGTGTGCTC; and round 2 forward primer: AATGATACGG CGACCACCGAGATCTACACGCCTCCCTCGCGCCATCAGAGATGTG.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Sequencing data can be obtained from the Gene Expression Omnibus (GEO) database using accession number GSE131506. All other data are available from the corresponding author upon reasonable request.

Code availability

The following programs were used. For sequence alignment, Bowtie2 2.3.4.1. For code development, Python v. 3.6.7. For read trimming, TrimGalore 0.4.1. For read quality assessment, FastQC v.0.11.8. For RNA secondary structure analysis, RNAstructure v.6.0.1. For calculating post-mapping statistics, Picard 2.18.7. For visualization of RNA secondary structure, VARNA v.3.93. For HIV-1 splicing analysis, https://github.com/SwanstromLab/SPLICING. For generating splice plots, R version 3.5.1. For figure construction, Adobe Illustrator CC 2019. For data analysis, Microsoft Excel 2018. For plot generation, Plotly v.3.2.1. The DREEM clustering algorithm is available at https://codeocean.com/capsule/0380995/tree.

31. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
32. Darty, K., Denise, A. & Ponty, Y. VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975 (2009).
33. 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).
34. Lahm, H. W. & Stein, S. Characterization of recombinant human interleukin-2 with micromethods. *J. Chromatogr. A* **326**, 357–361 (1985).

Acknowledgements The following reagents were obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: human recombinant IL-2 from M. Gately and HIV-1_{NL4-3} infectious molecular clone (pNL4-3) from M. Martin (cat. no. 114). This work was supported in part by the NIH (R21AI134365), the Center of HIV-1 RNA Studies (CRNA) NIH U54AI50470, the Smith Family Foundation and the Burroughs Wellcome fund.

Author contributions V.D.A.C., H.S., M.G., S.P., M.D.E., L.M., A.T.P. and S.R. developed and wrote the DREEM clustering algorithm and analysed validation studies. P.J.T. performed all cell and virus RNA modification assays. P.G. performed all in vitro RNA modification assays. P.J.T., P.G. and S.R. analysed HIV-1 RRE and A3 RNA structure data. P.J.T., S.R., T.Z. and P.B. designed mutants. T.Z. produced mutant plasmids. A.E. and R.S. performed splicing analysis assays. P.J.T. generated the genome-wide DMS-MaPseq library. P.J.T., H.S., P.G. and S.R. analysed genome-wide library data. T.C.T.L. conducted the U4 and U6 experiment. P.J.T. and S.R. wrote the manuscript. P.J.T., P.G. and S.R. created the figures. P.J.T., V.D.A.C., P.G., H.S., A.T.P., R.S., P.B., D.R.K., A.T. and S.R. edited the manuscript and figures.

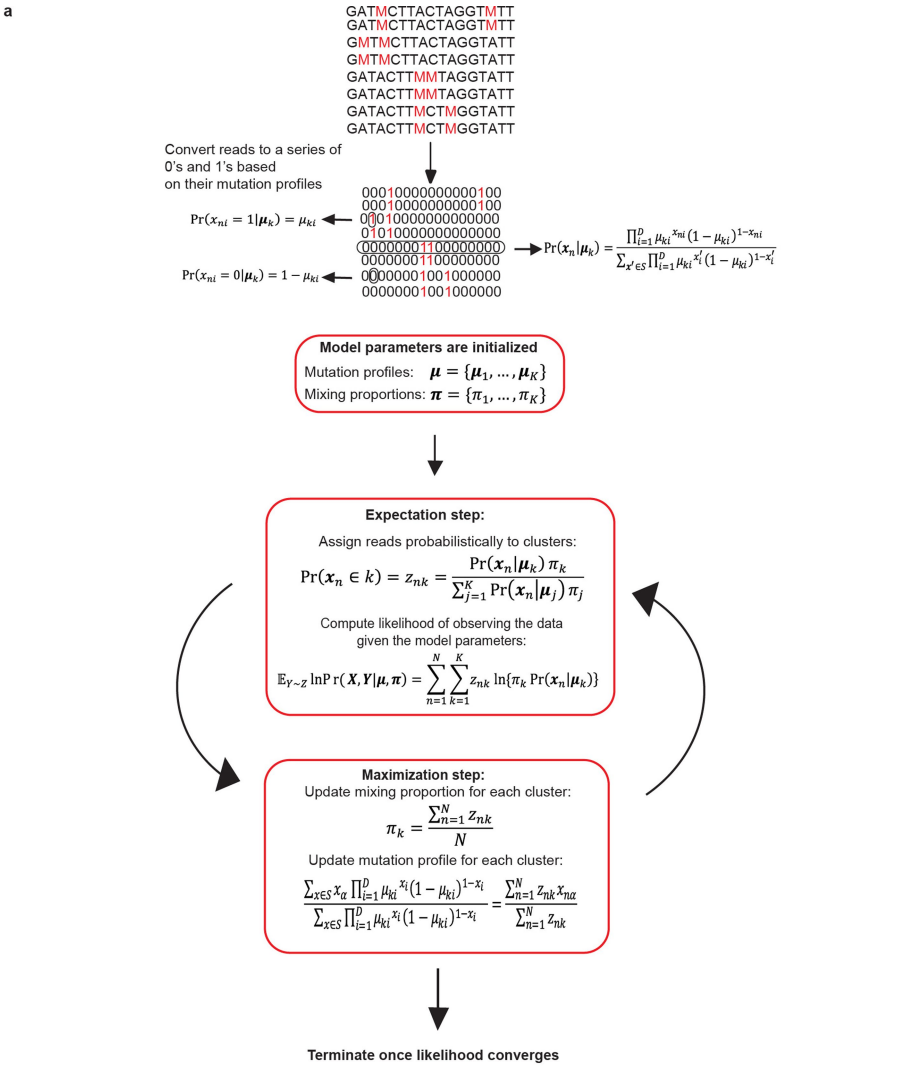
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2253-5.

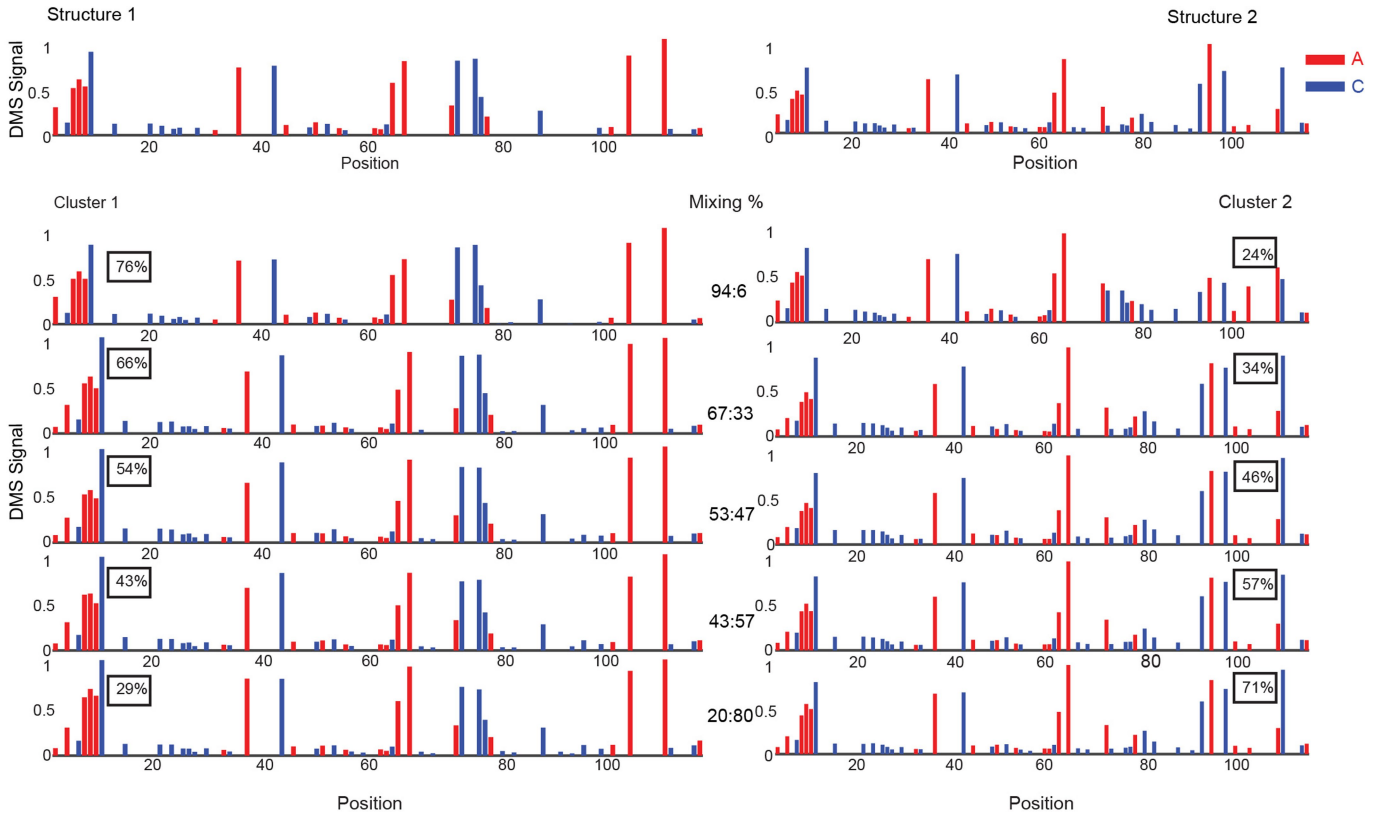
Correspondence and requests for materials should be addressed to S.R.

Peer review information *Nature* thanks Alan Frankel, Daniel Herschlag, Alain Laederach and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

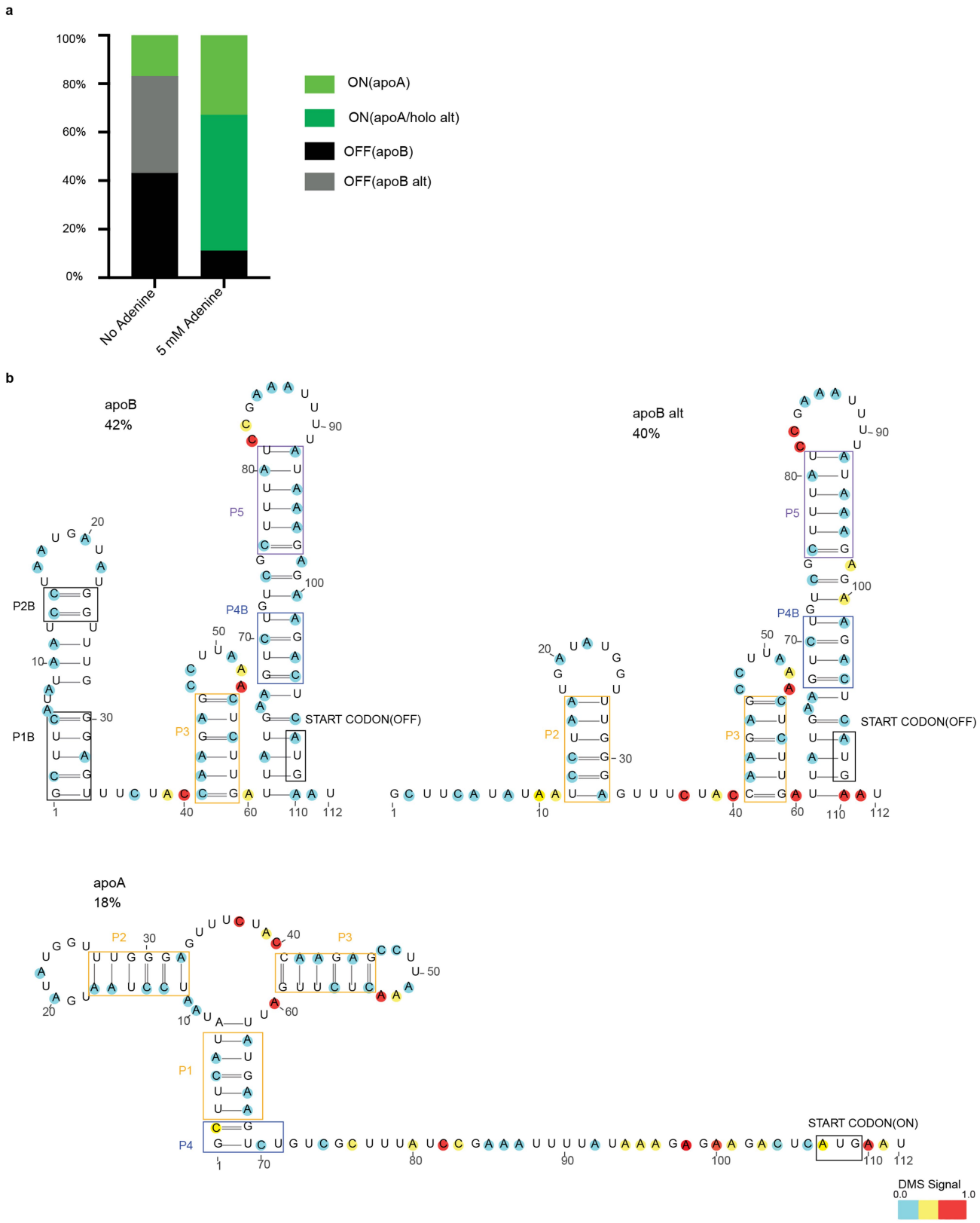


Extended Data Fig. 1 | See next page for caption.

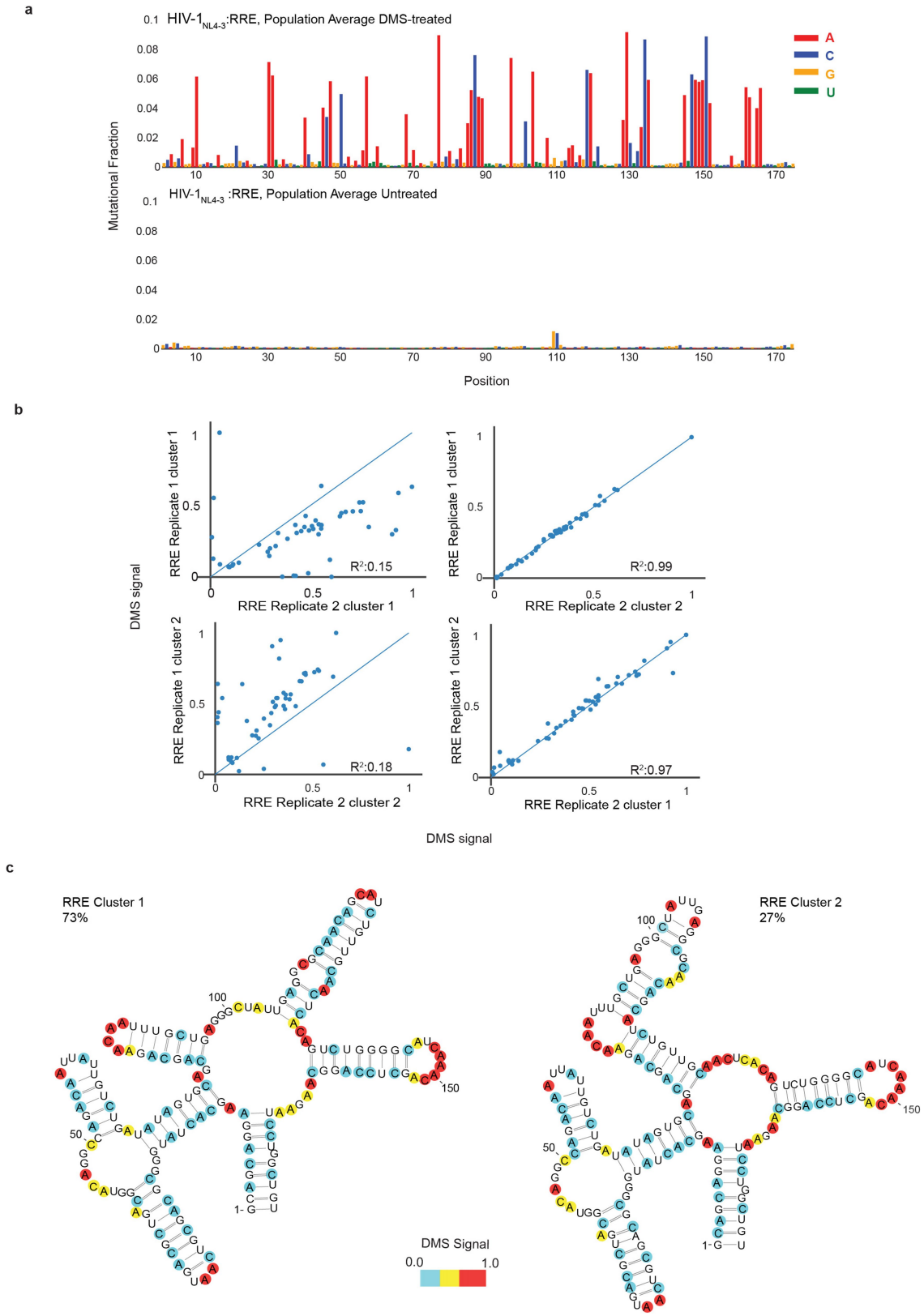
Extended Data Fig. 1 | DREEM clustering pipeline for DMS-MaPseq data. **a**, A read $\mathbf{x}''$ is represented as a series of D bits, in which D is the length of the read. A base is denoted by the bit 1 if it is mutated away from the reference, and by 0 otherwise. K is the number of clusters in the sample. $\boldsymbol{\mu}_k = \{\mu_{k1}, \mu_{ki}, \ldots, \mu_{kD}\}$ is the mutation profile of cluster k, and π_k is the mixing proportion of cluster k such that $\sum_{k=1}^{K} \pi_k = 1$ for $k = 1$ to K. The model parameters $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$ are randomly initialized. In the expectation step, reads are assigned probabilistically to clusters and the likelihood of observing the data given the model parameters is computed. In the maximization step, the mixing proportion is calculated from the read assignments and the mutation profiles are updated for each cluster to maximize the expectation value of the complete case likelihood. The expectation steps alternate with the maximization steps until the likelihood converges. The likelihood function is derived using Bernoulli mixture models modified to account for missing data in the form of the underrepresentation of reads with adjacent mutations. **b**, Mutational distance distribution between bases in denatured DMS-modified total RNA. The mutation distance versus frequency is plotted, between two DMS-reactive positions (that is, A or C to A or C; shown as yellow bars) and between one DMS-reactive position and a background mutation (for example, mutation owing to sequencing error) (that is, A or C to T or G; shown as blue bars). The blue bars demonstrate the frequency of observing two mutations due to background.



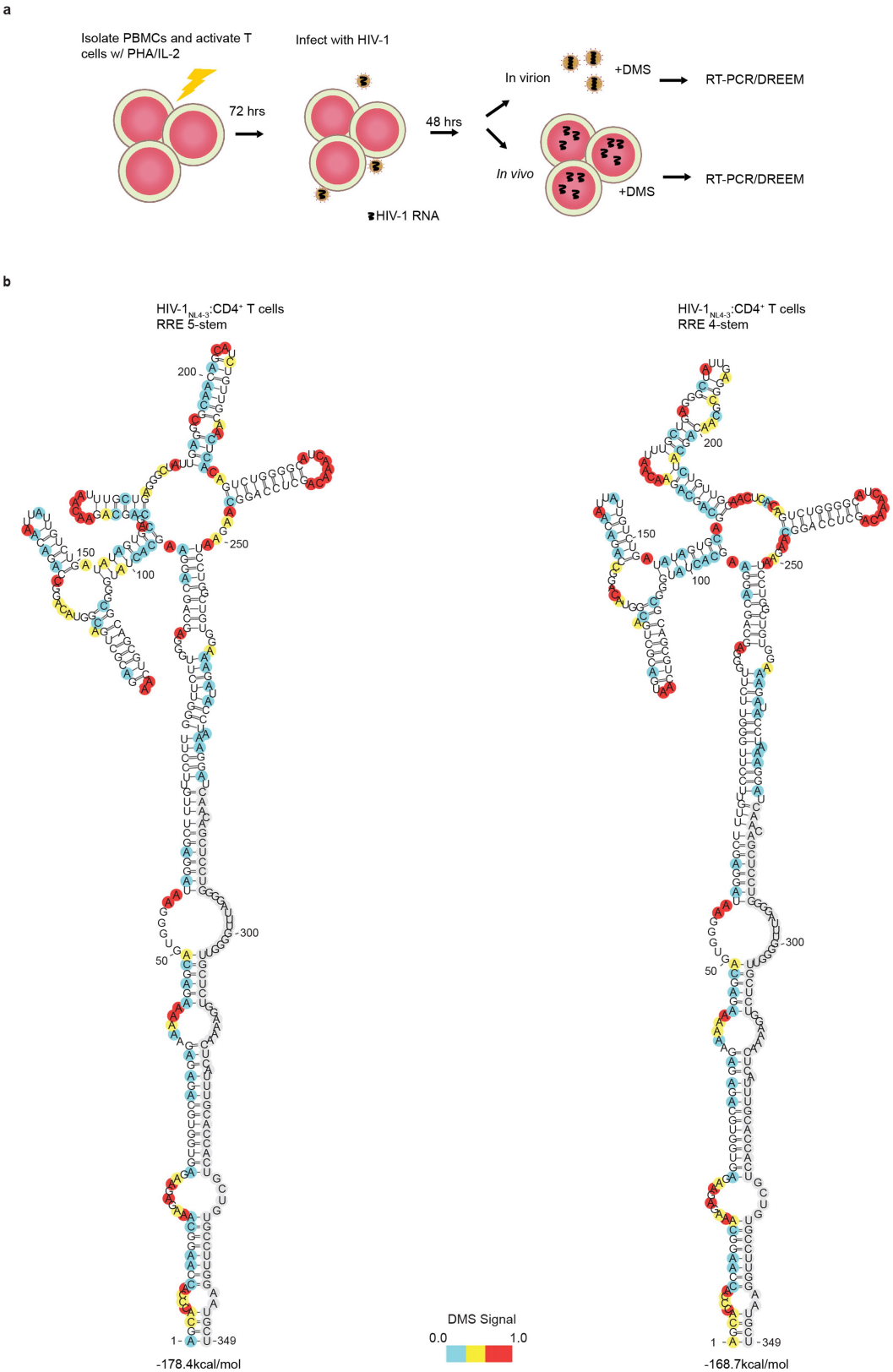
Extended Data Fig. 2 | DREEM clustering identifies and quantifies individual structures from in vitro mixing experiments. Structure 1 and structure 2 sequences were in vitro-transcribed and refolded, mixed in different proportions and probed with DMS-MaPseq. The region used for DREEM clustering covers nucleotides 21–135 (labelled as 1–115 on the figure), which excludes the primers used for RT–PCR (that have no DMS-induced mutations) and is identical in sequence for the two structures, except for the A > C mutation at position 94. Position 94 is masked during analysis. The topmost panel shows the DMS reactivity pattern of structure 1 by itself and structure 2 by itself. The rest of the panels show the clustering results at specified mixing ratios ($n = 1$).



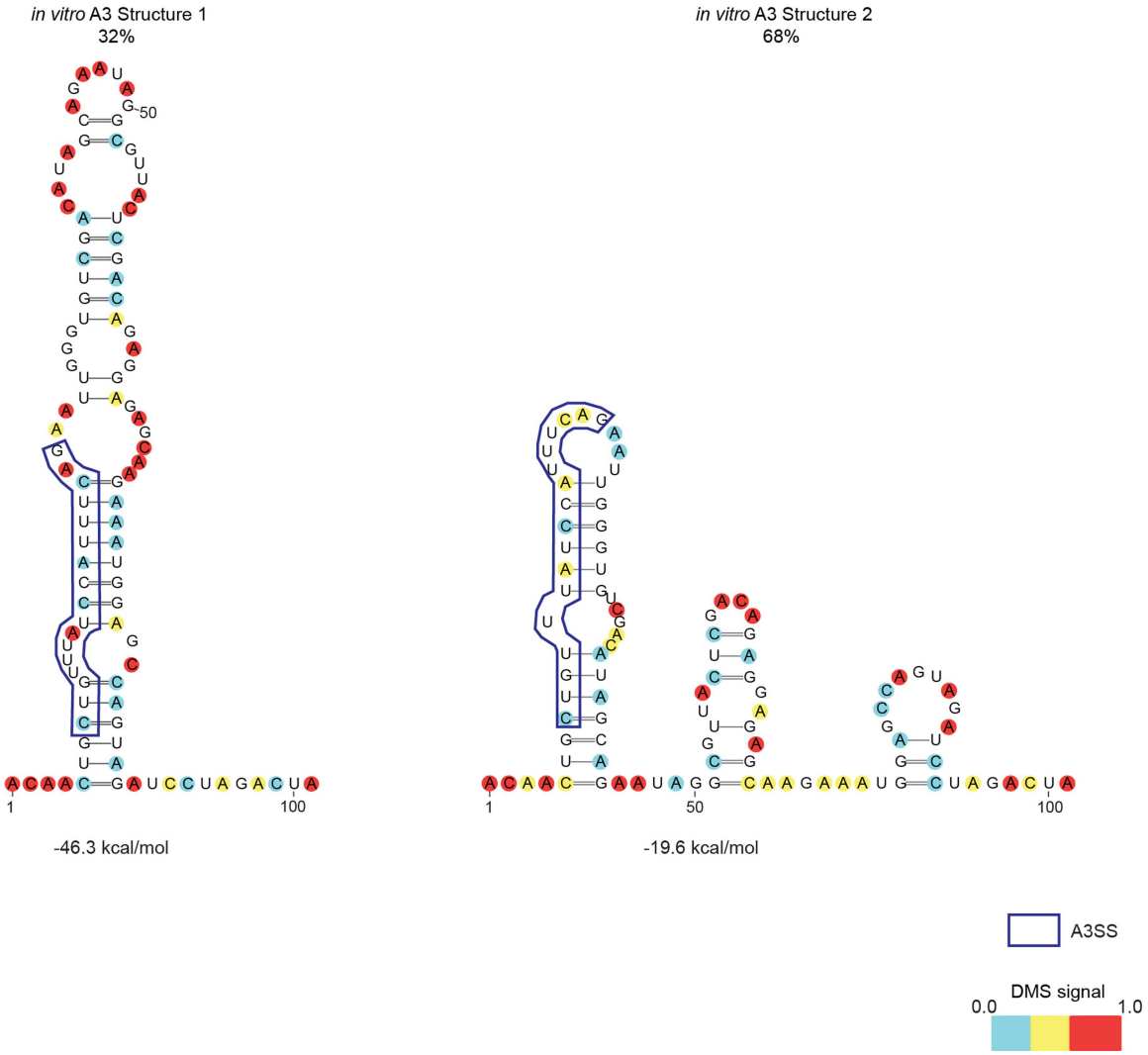
Extended Data Fig. 3 | Secondary structure models for the *V. vulnificus add* riboswitch. a, Percentages for each cluster detected in the presence or absence of 5 mM adenine to the *add* riboswitch. **b**, In vitro structure models obtained from probing *add* using DMS-MaPseq followed by DREEM, colour-coded by normalized DMS signal. The ApoB and ApoB alternative structures represent the off state, which is incompetent for ligand binding. ApoA represents the on state. Previously identified helices are boxed and labelled.



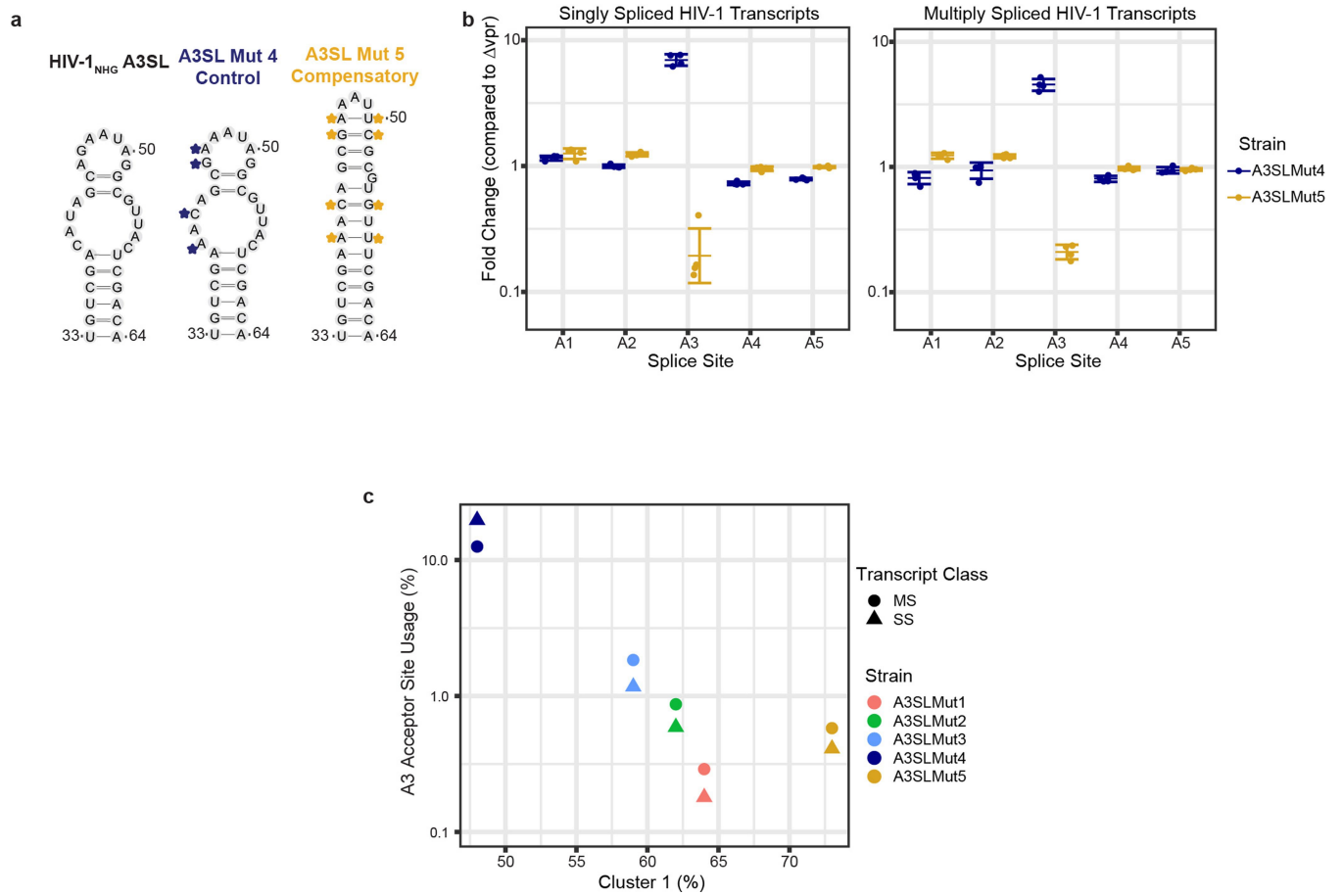
Extended Data Fig. 4 | DREEM clustering reveals an equilibrium of four-stem and five-stem structures for the in vitro-folded HIV-1 RRE. **a**, Population-average DMS-MaPseq data for in vitro-transcribed, refolded and DMS-treated (or untreated) samples. **b**, Scatter plots showing the reproducibility of the DMS signal from the DREEM clustering results between two replicates with different DMS modification conditions. Replicate 1 was modified in 0.25% DMS and replicate 2 was modified with 2.5% DMS. R^2 is Pearson's R^2. **c**, DREEM clustering data from **b** were used as constraints to generate RNA structure models. The models derived for clusters 1 and 2 from replicate 1 are shown, colour-coded by normalized DMS signal.



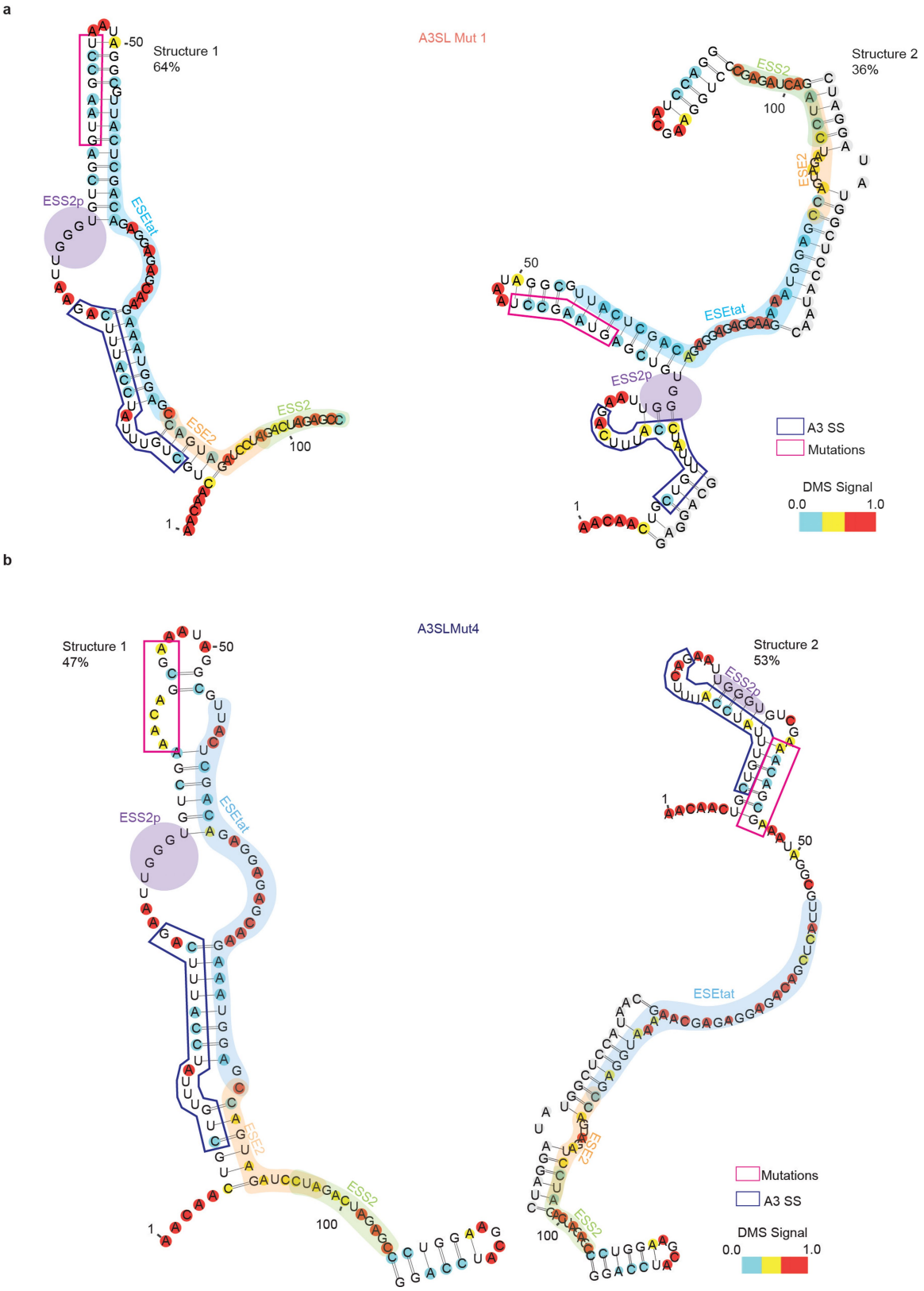
Extended Data Fig. 5 | The HIV-1 RRE forms two stable alternative structures in CD4$^+$ T cells. a, Schematic of DMS treatment in primary cells and isolated virions. **b**, DMS-MaPseq probing of the intracellular HIV-1_{NL4-3} RRE in CD4$^+$ T cells was used as input for DREEM clustering. Two clusters passed the BIC test and were used as constraints on the folding using RNAstructure. Structural models are colour-coded by normalized DMS reactivity; bases not covered by the region of PCR are coloured in grey. Data used to construct models are representative data from $n = 2$ biologically independent experiments.



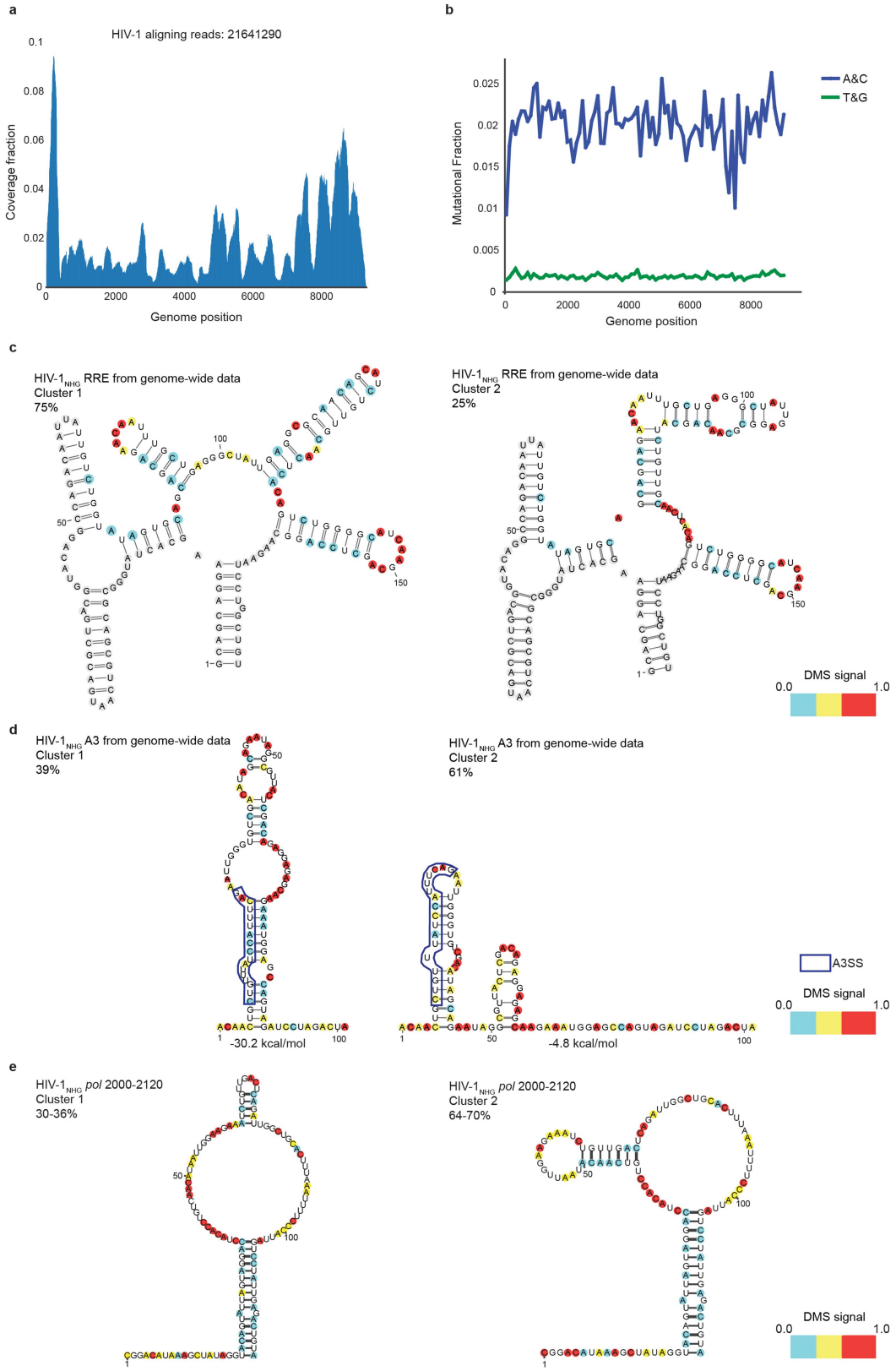
Extended Data Fig. 6 | The A3 splice site forms alternative structures in vitro. A 472-nt A3 sequence from the HIV-1_{NHG} strain was in vitro-transcribed, refolded and probed with DMS-MaPseq. Models based on DREEM clustering for the local structures that form at the A3 site are shown, colour-coded by the normalized DMS signal. Percentages of clusters 1 and 2 come from $n = 1$ experiment, as determined by DREEM.



Extended Data Fig. 7 | Splice site use in additional A3 mutants. a, Structure models illustrating the mutant design for A3SL mut4 and A3SL mut 5. **b**, Splice site use for A3SL mut 4 and A3SL mut 5 for splice sites A1–A5, reported as fold change compared to Δ*vpr* HIV-1_{NHG}. Central bar represents the mean, and error bars indicate s.d. $n = 4$ biologically independent experiments. **c**, Average fraction of transcripts using the A3 site, compared to the percentage of cluster 1 (A3SL), as determined by DREEM ($n = 1$) for A3SL mut 1–5. Mutants are colour-coded. A dot indicates a multiply spliced (MS) HIV-1 transcript, and a triangle indicates a singly splice (SS) HIV-1 transcript.

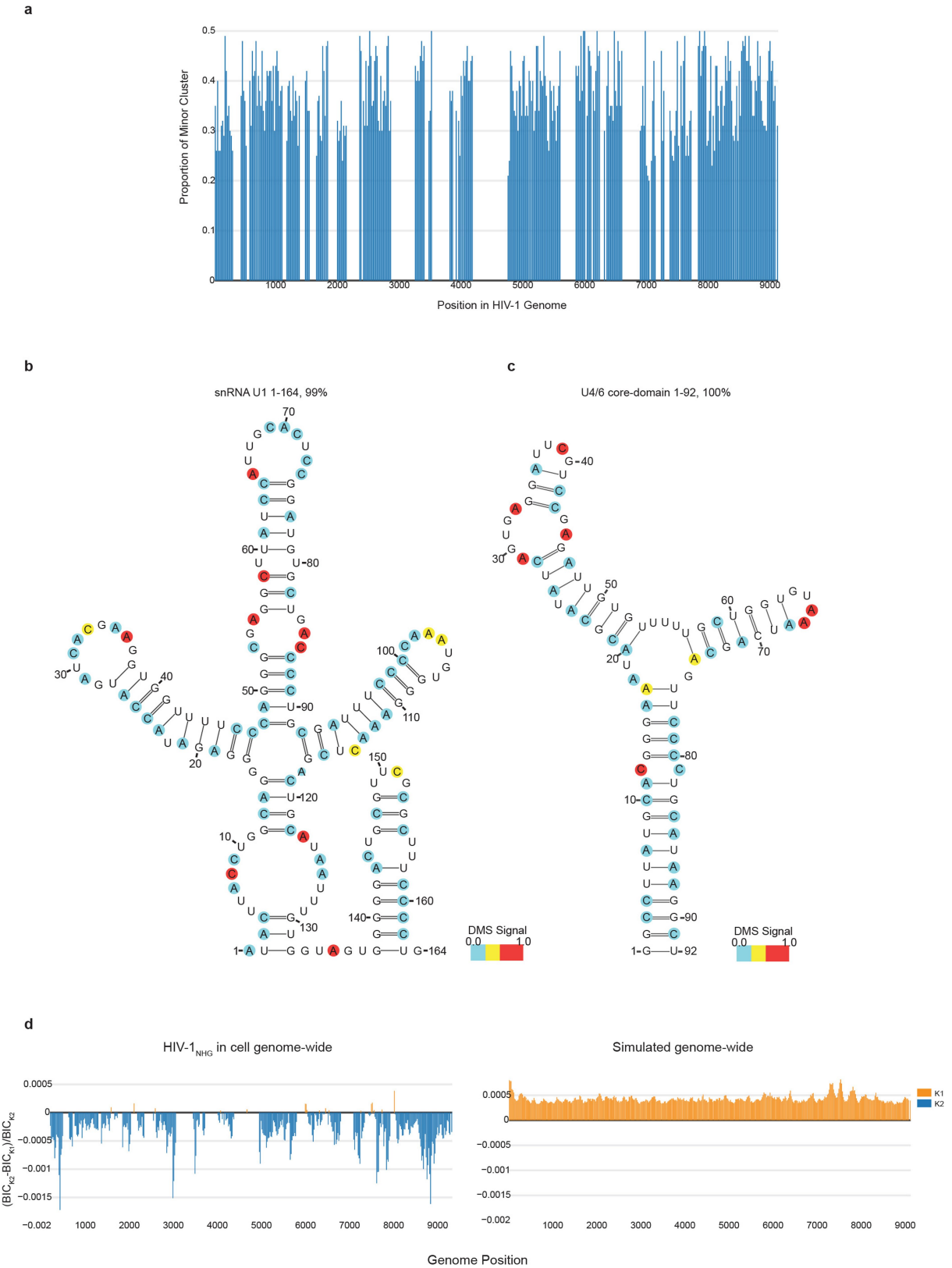


Extended Data Fig. 8 | Structural models of A3SL mut 1 and A3SL mut 4. **a**, Structural models for A3SL mut 1 derived from $n = 1$ experiment, after DREEM clustering. Pink box, region of mutations; blue box, splice site. Exonic splicing enhancer (ESE) and exonic splicing silencer (ESS) binding sites are shown. **b**, Structural models made using DMS-MaPseq data from HEK293T cells transfected with Δ*vpr* HIV-1_{NHG} A3SL mut 4. Dark blue box, sequence of the A3 splice site; pink box, location of the mutations. Splice enhancer and suppressor binding sites are highlighted; purple, ESS2p; blue, ESEtat; orange, ESE2; and green, ESS2. Percentages of each cluster come from $n = 1$ experiment.

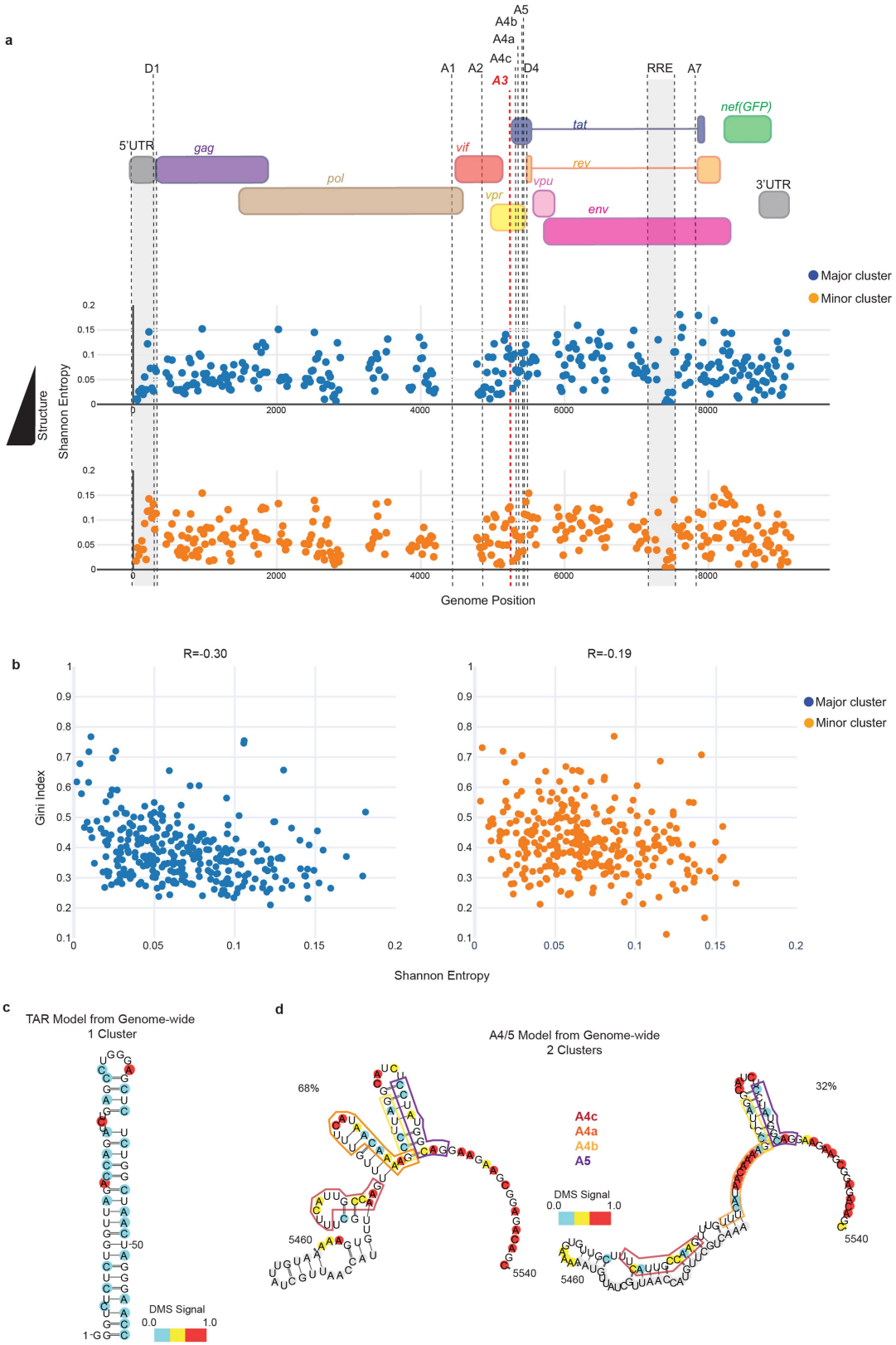


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Quality control for the generation of the genome-wide HIV-1_{NHG} library. a, Coverage of HIV-1 genome with DMS-MaPseq data from HEK293T cells transfected with HIV-1_{NHG}. **b**, Moving average of A and C mutation frequency in 100-nt windows after DMS-MaPseq, compared to the moving average T and G mutation frequency. **c**, DMS-MaPseq data from HEK293T cells transfected with HIV-1_{NHG} were used as input for DREEM. Local 80-nt window from Fig. 4 for the RRE region was used for clustering. Percentages of clusters 1 and 2 come from $n = 1$ experiment. Nucleotides are colour-coded on the basis of the normalized DMS signal; bases outside of the window used for clustering are coloured in grey. **d**, The A3 splice site was analysed using DMS-MaPseq and DREEM clustering from genome-wide data from HEK293T cells transfected with HIV-1_{NHG}. Percentages of clusters 1 and 2 come from $n = 1$ experiment, as determined by DREEM. Nucleotides are colour-coded on the basis of the normalized DMS signal. **e**, A region of the HIV-1 genome in the *pol* coding region (nucleotides 2,000–2,120, based on HIV-1_{NHG} genomic RNA coordinates) was analysed using DMS-MaPseq and DREEM clustering from genome-wide data from HEK293T cells transfected with HIV-1_{NHG}. Two clusters passed the BIC test in adjacent 80-nt windows that overlapped by 40 nt. The two 80-nt windows were combined to make the structural models. The range of proportions of each cluster come from the individual windows of $n = 1$ experiment. Nucleotides were colour-coded on the basis of the normalized DMS signal.



Extended Data Fig. 10 | Proportion of minor clusters across the HIV-1 genome and the U1, U4 and U6 core-domain structural models. a, Each bar shows the proportion of a minor cluster of an 80-nt window as a function of genome position for regions in the HIV-1_{NHG} genome dataset that are covered by at least 100,000 reads and pass 2 clusters, according to the BIC test. **b**, U1 structural prediction from HEK293T cells transfected with HIV-1_{NHG}. The abundance of the cluster was obtained from DREEM clustering. **c**, In vitro DMS-modified U4 and U6 core-domain RNA. The structure is shown for a population average; cluster 2 did not pass the BIC test. **d**, Left, difference in BIC test value between $K = 2$ and $K = 1$, normalized to the value for $K = 2$ for the real whole-genome dataset. Each bar represents an 80-nt window across the HIV-1 genome. Orange, windows in which only one cluster was detected according to the BIC test; blue, clusters for which two clusters passed the BIC test. Right, the same plot as shown in the left panel from simulated data, for which the mutations were randomly distributed but had the same average number of mutations per read as the true data.



Extended Data Fig. 11 | Shannon entropy across the HIV-1 genome, and A4 and A5 splice sites. a, Overlay of the HIV-1_{NHG} genomic organization on top of a Shannon entropy plot. Each dot represents an 80-nt window, in which Shannon entropy was calculated from DMS reactivity. The top plot is the major cluster and the bottom is the minor cluster. **b**, Scatter plot of Gini index versus Shannon entropy for the major and minor clusters ($n = 1$). R^2 is Pearson's R^2. **c**, Structural model of the transcription-activation-region stem loop from the genome-wide DMS-MaPseq and DREEM data. **d**, Structural model from two clusters found using the genome-wide DMS-MaPseq and DREEM data for a window containing four splice acceptor sites (A4a, A4b, A4c and A5). Splice sites are boxed. Nucleotides are colour-coded on the basis of the normalized DMS signal.

nature research

Corresponding author(s): Silvi Rouskin

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No collection software used
Data analysis	Sequence alignment: Bowtie2 2.3.4.1. For code development: python v. 3.6.7. For read trimming: TrimGalore 0.4.1. For read quality assessment: FastQC v0.11.8. For RNA secondary structure analysis: RNAstructure v6.0.1. For calculating post-mapping statistics: Picard 2.18.7. RNA secondary structure visualization: VARNA v3.93. HIV-1 splicing analysis: https://github.com/SwanstromLab/SPLICING. Splice plot creation: R version 3.5.1. For figure construction: Adobe Illustrator CC 2019. For data analysis: Microsoft Excel 2018. Plot generation: Plotly v3.2.1. DREEM clustering algorithm is available at https://codeocean.com/capsule/0380995/tree Sfold 2.2 http://sfold.wadsworth.org/cgi-bin/index.pl

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data available through GEO accession code GSE 131506.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For clustering analysis, we used a minimum sequencing depth of 100,000 reads per base.
Data exclusions	DMS-modified samples that did not have sufficient signal (mutational fraction <0.05 A/C for population average) were excluded from analysis.
Replication	Reproducibility of clustering results is shown in figures 1, 2, and 3, as well as extended data figure 4.
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293t cells obtained from ATCC.
Authentication	Cell line was not authenticated.
Mycoplasma contamination	Cells tested negative for mycoplasma following manufacturer's specifications for the LookOut kit from Millipore-Sigma.
Commonly misidentified lines (See ICLAC register)	N/A