

A General Computational Framework for Sequence-Level Thermodynamic Analysis of Modeled RNA–DNA and DNA–DNA Duplex States: A Proof-of-Concept Study in HIV-1

SAADAT SAMADI DINANI

saada.1349@gmail.com

Independent Researcher <https://orcid.org/0009-0008-2861-7471>

Research Article

Keywords: HIV-1, RNA-DNA hybrid, R-loop, thermodynamics, RNAstructure, computational framework, model-defined thermodynamic contrast, sequence-level analysis

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11394670/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

A General Computational Framework for Sequence-Level Thermodynamic Analysis of Modeled RNA–DNA and DNA–DNA Duplex States: A Proof-of-Concept Study in HIV-1

Abstract

We present a sequence-level computational framework designed to be applicable across viral genomes and demonstrate it in a proof-of-concept analysis of HIV-1. Sixteen predefined regions of the HXB2 reference genome (K03455.1) were evaluated using two RNAstructure 6.6 procedures: OligoWalk for an RNA-target/DNA-oligo duplex calculation and DuplexFold for a DNA–DNA bimolecular minimum-free-energy reference. We define a model-specific contrast, $\Delta\Delta G_{\text{model}} = \Delta G_{\text{RD}} - \Delta G_{\text{DD}}$, rather than a universal same-algorithm $\Delta\Delta G$. Fifteen of 16 regions had negative contrasts; the panel mean was -8.00 kcal/mol (SD 5.45). A two-sided one-sample test against zero gave $t_{15} = -5.87$, $p = 1.54 \times 10^{-5}$. The directional result persisted in overlap-sensitivity analyses excluding TAR and/or R_5LTR. These results constitute a sequence-level, testable computational prediction under the specified modeling procedures. They do not establish cellular R-loop formation, genomic integration, or a causal biological mechanism. The study is limited to one reference genome and to 16 non-independent, partly overlapping regions.

Keywords: HIV-1; RNA–DNA hybrid; R-loop; thermodynamics; RNAstructure; computational framework; model-defined thermodynamic contrast

1 Introduction

R-loops are three-stranded nucleic-acid structures in which an RNA strand forms a hybrid with one DNA strand while the other DNA strand is displaced. They are established features of genome biology and can be associated with transcription, genome instability, and regulatory processes [1–4]. Viral genomes provide a useful setting in which sequence composition and predicted duplex energetics can be examined systematically, but sequence-level thermodynamic calculations should not be conflated with direct evidence of cellular R-loop formation [6–8].

HIV-1 provides a suitable proof-of-concept system because a complete reference genome is available and individual functional regions can be defined reproducibly. The present study uses the HXB2 reference sequence K03455.1 [9–11, 32]. Recent experimental work has reported HIV-1-associated host genomic R-loops [31]. Such observations motivate computational prioritization of viral sequence contexts, but they do not make a sequence-only thermodynamic calculation equivalent to a cellular R-loop measurement. Integration-site selection in HIV-1 is governed by multiple molecular and chromatin-level factors [12–15].

The methodological objective here is deliberately narrower. We establish a reproducible pipeline that (i) extracts predefined viral regions from a locked reference genome, (ii) evaluates an RNA–DNA modeled duplex state, (iii) evaluates a DNA–DNA reference state, (iv) calculates

a study-specific thermodynamic contrast, and (v) quantifies the direction and magnitude of that contrast across the selected panel. The framework is designed for application to other viral genomes, whereas the present HIV-1 analysis is a single-genome proof of concept.

1.1 Study question and scope

The primary question is whether the modeled RNA–DNA quantity is systematically more negative than the corresponding DNA–DNA reference quantity across the 16 predefined HIV-1 regions under the stated computational procedures. A negative value of the contrast means only that the RNA–DNA value is more negative than the DNA–DNA reference in this model comparison. It is not, by itself, evidence that a cellular R-loop forms.

2 Methods

2.1 Reference genome and predefined regions

The complete HIV-1 HXB2 reference genome K03455.1 was used as the locked sequence source. The reference genome is 9719 nt in length [32]. Coordinates were treated as 1-based and inclusive. The 16 predefined regions are listed in Table 1. TAR and R_5LTR overlap by design; this overlap was retained in the primary analysis and explicitly tested by sensitivity analysis.

Table 1: Sixteen predefined HIV-1 regions used in the proof-of-concept analysis. Coordinates are 1-based and inclusive.

Region	Coordinates	Length (nt)
TAR	454–482	29
R_5LTR	454–551	98
PBS	182–199	18
U5	552–634	83
gag	790–889	100
pol	2085–2184	100
env	6225–6324	100
nef	8797–8896	100
RRE	7710–7809	100
vif	5041–5140	100
vpr	5559–5658	100
vpu	6062–6161	100
rev	8379–8478	100
ins	8965–8992	28
asp	9108–9207	100
V1V2	6581–6712	132

For each region, the reference DNA sequence was extracted directly from K03455.1. An RNA representation was generated by converting T to U without changing the sequence order. For the DNA–DNA reference, the complementary strand was generated as the reverse complement of the extracted DNA sequence.

2.2 Thermodynamic procedures

RNAstructure provides computational tools for nucleic-acid structure and thermodynamic analysis, including bimolecular calculations and OligoWalk [16–19]. The production RNA–DNA observable was obtained with OligoWalk using an RNA target and a DNA oligo. The `-d` option specifies DNA parameters for the oligo, while `-m 3` specifies that target structure is not considered in the OligoWalk calculation. The reported `Duplex` value was retained as the RNA–DNA quantity.

We denote this quantity as

$$\Delta G_{\text{RD}} = \text{OligoWalk Duplex value (RNA target + DNA oligo, -d, -m 3)}. \quad (1)$$

For the DNA–DNA reference, DuplexFold was used with DNA parameters (`-d`) for the two strands. The reported bimolecular minimum free energy was retained:

$$\Delta G_{\text{DD}} = \text{DuplexFold bimolecular MFE (DNA + reverse-complement DNA, -d)}. \quad (2)$$

The primary study-specific contrast was then defined as

$$\Delta\Delta G_{\text{model}} = \Delta G_{\text{RD}} - \Delta G_{\text{DD}}. \quad (3)$$

A negative $\Delta\Delta G_{\text{model}}$ therefore indicates that the modeled RNA–DNA quantity is more negative than the corresponding DNA–DNA reference quantity under the two specified computational procedures.

Because the two observables are produced by different RNAstructure calculations, this quantity is explicitly treated as a *model-defined thermodynamic contrast*, not as a universal same-algorithm $\Delta\Delta G$ [20].

The exact command-line behavior of OligoWalk and DuplexFold was taken from the RNAstructure 6.6 documentation. The calculations used the program defaults for temperature and thermodynamic parameters; the production runs corresponded to 37 °C.

2.3 Production values

The authoritative production values are given in Table 2.

Table 2: Production thermodynamic values for the 16 predefined HIV-1 regions. All energies are in kcal/mol.

Region	Length	ΔG_{RD}	ΔG_{DD}	$\Delta\Delta G_{model}$	$\Delta\Delta G/nt$
TAR	29	-39.4	-37.8	-1.6	-0.0552
R_5LTR	98	-142.8	-134.8	-8.0	-0.0816
PBS	18	-22.9	-22.0	-0.9	-0.0500
U5	83	-108.8	-109.7	+0.9	+0.0108
gag	100	-140.1	-127.9	-12.2	-0.1220
pol	100	-145.6	-137.5	-8.1	-0.0810
env	100	-147.0	-133.9	-13.1	-0.1310
nef	100	-155.5	-139.2	-16.3	-0.1630
RRE	100	-151.1	-138.8	-12.3	-0.1230
vif	100	-133.2	-126.3	-6.9	-0.0690
vpr	100	-145.5	-134.2	-11.3	-0.1130
vpu	100	-117.8	-115.5	-2.3	-0.0230
rev	100	-156.4	-140.1	-16.3	-0.1630
ins	28	-44.9	-37.8	-7.1	-0.2536
asp	100	-142.4	-132.7	-9.7	-0.0970
V1V2	132	-162.9	-160.1	-2.8	-0.0212

2.4 Statistical analysis

All statistical calculations were performed with Python 3.8.10 using NumPy, SciPy, and pandas. The primary null hypothesis was

$$H_0 : \mu_{\Delta\Delta G} = 0, \quad (4)$$

with a two-sided one-sample t -test used as the primary inferential test.

The analysis was intentionally limited in scope because the 16 regions are measurements from a single reference genome and are not independent biological replicates.

Directional and distributional sensitivities included an exact sign test and a sign-flip permutation test with 10,000 permutations and seed 42.

Pearson and Spearman correlations with sequence length and GC content were treated as exploratory. No post-hoc multivariable model was fitted because $n = 16$ is small and the regions are not independent.

Overlap sensitivity was assessed by repeating the primary mean analysis after excluding TAR, excluding R_5LTR, and excluding both overlapping regions.

Bootstrap confidence intervals for the panel mean used 100,000 resamples with seed 20261004.

These intervals describe uncertainty in the observed region panel and should not be interpreted as confidence intervals across biological replicates.

2.5 Reproducibility and quality control

The sequence extraction, orientation, OligoWalk, DuplexFold, and arithmetic reconstruction audits were completed before statistical analysis.

The primary production pipeline was restricted to the locked K03455.1 reference and the specified RNAstructure 6.6 procedures.

Earlier exploratory scripts that converted RNA to DNA and used incompatible calculations were excluded from the production pipeline.

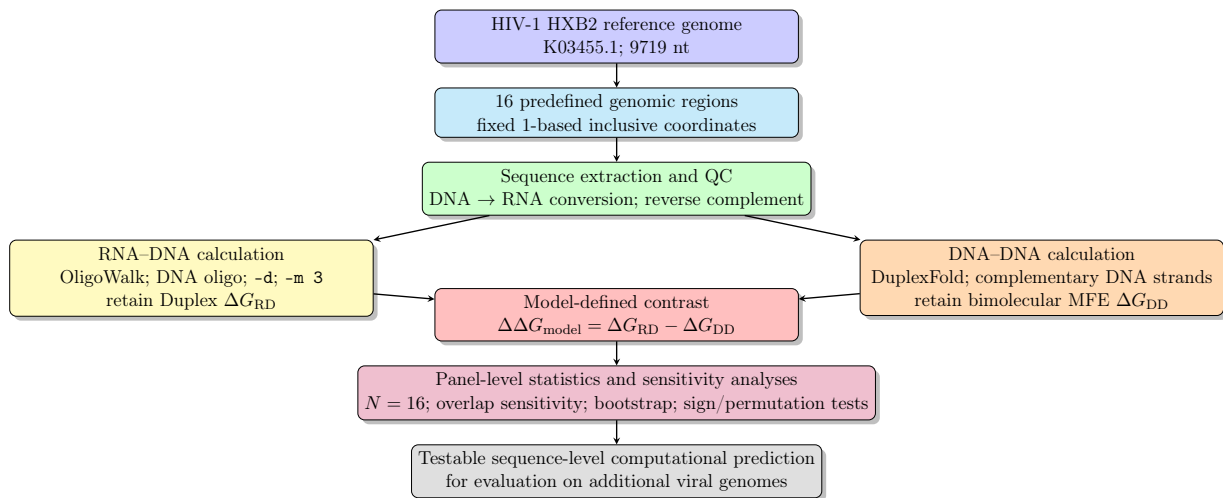


Figure 1: Computational workflow used in the proof-of-concept study. The locked HIV-1 K03455.1 reference is used to extract 16 predefined regions. Each region is evaluated as an RNA–DNA modeled duplex with OligoWalk and as a DNA–DNA reference duplex with DuplexFold. The resulting model-defined contrast is summarized statistically and subjected to overlap and exploratory sequence-composition analyses. The workflow produces a testable computational prediction; it is not direct evidence of cellular R-loop formation or genomic integration.

3 Results

3.1 Primary thermodynamic pattern

Across the 16 predefined HIV-1 regions, 15 had negative $\Delta\Delta G_{\text{model}}$ values and one had a positive value.

The mean contrast was -8.00 kcal/mol (SD 5.45), with median -8.05 kcal/mol, minimum -16.30 kcal/mol, and maximum $+0.90$ kcal/mol.

The primary two-sided one-sample test against zero gave

$$t_{15} = -5.87, \quad p = 1.54 \times 10^{-5}.$$

The most negative contrasts were observed for nef and rev (both -16.3 kcal/mol), followed by env (-13.1 kcal/mol), RRE (-12.3 kcal/mol), and gag (-12.2 kcal/mol).

U5 was the only positive region, with a contrast of $+0.9$ kcal/mol.

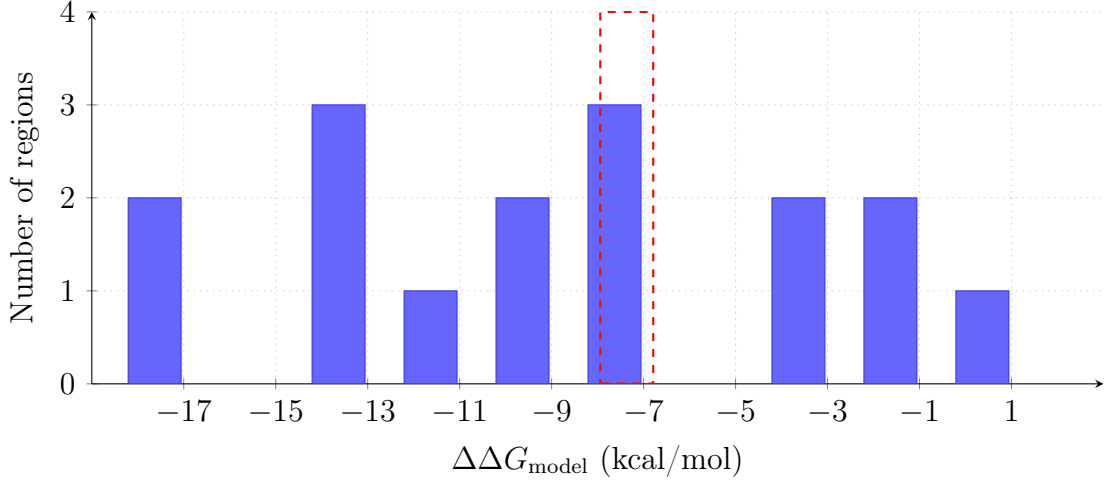


Figure 2: Distribution of the model-defined thermodynamic contrast $\Delta\Delta G_{\text{model}}$ across the 16 predefined HIV-1 regions. The histogram summarizes the frequency of regions in each 2 kcal/mol energy bin centered at the indicated tick positions. The dashed red vertical line indicates the panel mean (-8.00 kcal/mol).

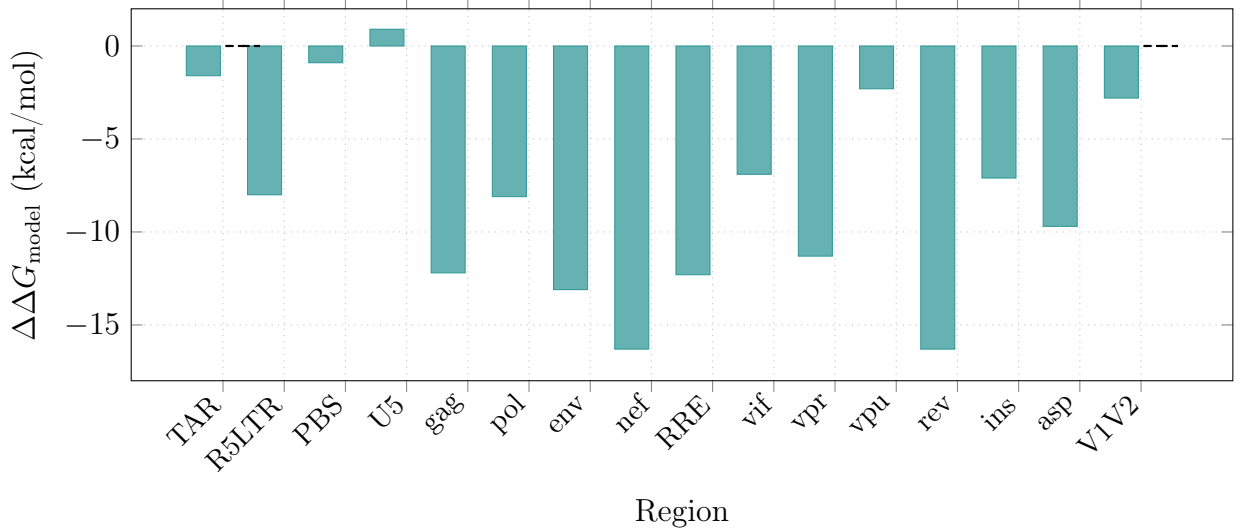


Figure 3: Regional model-defined thermodynamic contrasts for the 16 predefined HIV-1 regions. Negative values indicate that the calculated RNA–DNA duplex quantity was more negative than the corresponding DNA–DNA reference quantity under the specified computational procedures.

3.2 Overlap sensitivity

The directional pattern remained stable when the overlapping regions were removed. Excluding TAR left 14 negative regions among 15. Excluding R_5LTR also left 14 of 15 negative. Excluding both left 13 of 14 negative. The corresponding means were -8.43 , -8.00 , and -8.46 kcal/mol, respectively.

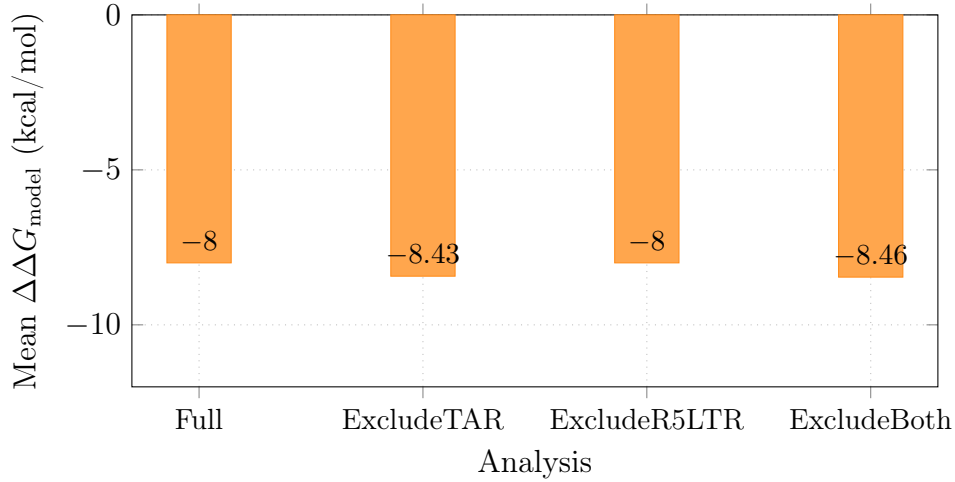


Figure 4: Overlap-sensitivity analysis. The panel-level mean model-defined thermodynamic contrast remains negative after exclusion of TAR, R_5LTR, or both overlapping regions.

For the full panel, the bootstrap 95% confidence interval for the mean was $[-10.55, -5.39]$ kcal/mol. After excluding both overlapping regions, the corresponding interval was $[-11.21, -5.61]$ kcal/mol.

The exact sign test for 15 negative values among 16 gave $p = 2.59 \times 10^{-4}$, and the sign-flip permutation test gave $p = 9.999 \times 10^{-5}$.

3.3 Exploratory sequence associations

Pearson correlation between raw $\Delta\Delta G_{\text{model}}$ and region length was $r = -0.414$, $p = 0.111$. The corresponding Spearman correlation was $\rho = -0.501$, $p = 0.048$. For normalized $\Delta\Delta G_{\text{model}}/\text{nt}$, Pearson correlation with length was $r = 0.166$, $p = 0.539$.

Raw contrast showed no statistically significant Pearson association with GC content: $r = -0.279$, $p = 0.295$. The normalized contrast showed an exploratory association with GC content: $r = -0.551$, $p = 0.027$, and $\rho = -0.538$, $p = 0.032$.

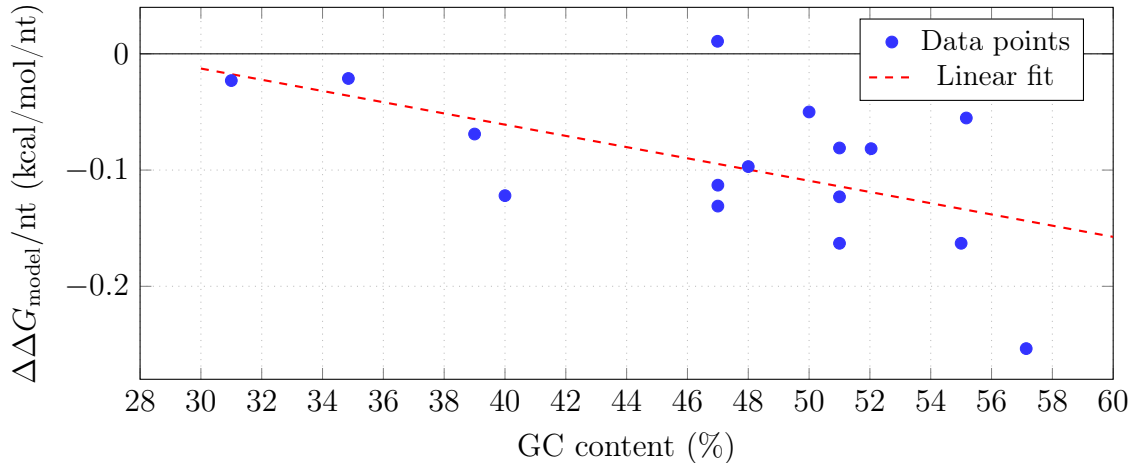


Figure 5: Exploratory relationship between GC content and normalized model-defined thermodynamic contrast. Each point represents one of the 16 predefined HIV-1 regions. The association is exploratory and is not used to modify the primary statistical inference.

3.4 Summary statistics

Table 3: Summary of the primary panel-level analysis.

Quantity	Value
Number of predefined regions	16
Negative contrasts	15/16 (93.75%)
Mean $\Delta\Delta G_{\text{model}}$	-8.00 kcal/mol
SD	5.45 kcal/mol
Median	-8.05 kcal/mol
One-sample t statistic	$t_{15} = -5.87$
Two-sided p value	1.54×10^{-5}
Exact sign-test p	2.59×10^{-4}
Sign-flip permutation p	9.999×10^{-5}
Bootstrap 95% CI for mean	$[-10.55, -5.39]$ kcal/mol

4 Discussion

4.1 Interpretation of the primary result

The central result is a reproducible sequence-level pattern: 15 of the 16 predefined regions yielded a negative model-defined thermodynamic contrast. In the terms of the present framework, the modeled RNA–DNA quantity was more negative than the DNA–DNA reference in those regions.

The result is statistically distinguishable from zero at the level of the selected region panel. This statement must be kept separate from a biological claim about R-loops.

An RNA–DNA duplex calculation does not model chromatin accessibility, transcriptional state, strand displacement, protein competition, local topology, cellular concentration, or kinetic formation pathways. Therefore, the present results are best interpreted as a computational prioritization signal and a testable sequence-level prediction.

4.2 Relationship to prior R-loop literature

R-loop formation is influenced by sequence, transcription, RNA processing, DNA topology, and protein factors [1, 4, 5, 8]. The present analysis addresses only a narrow thermodynamic component of that broader mechanism.

Park et al. reported HIV-1-associated host genomic R-loops experimentally [31]; however, the present work neither reproduces that cellular measurement nor assigns the modeled viral regions as integration hotspots [12, 13].

Instead, the framework can be used to identify sequence contexts that may merit controlled RNA–DNA hybridization experiments.

4.3 Why the model-defined contrast is useful

A raw RNA–DNA free energy alone does not provide the comparative information sought here. The study therefore uses a DNA–DNA reference for the same extracted sequence context. This contrast makes the computational question explicit.

Because OligoWalk and DuplexFold are different calculations, the contrast should not be interpreted as a universal thermodynamic constant. Its value is tied to the software, parameters, sequence construction, and command-line settings specified in the Methods.

4.4 Overlap and sequence composition

TAR and R_5LTR overlap, so their joint inclusion could inflate apparent consistency if treated as independent evidence. The overlap-sensitivity analysis showed that the negative direction persisted after excluding either region or both.

Length and GC content were examined as exploratory descriptors. Given the small, non-independent panel, these observations should be treated as hypothesis-generating rather than as adjusted causal effects.

4.5 Limitations

The most important limitations are:

1. The analysis uses one reference genome, K03455.1/HXB2.
2. The 16 regions are not independent biological replicates.
3. TAR and R_5LTR overlap.
4. The RNA–DNA and DNA–DNA observables are obtained by different computational procedures.
5. The model does not include chromatin accessibility, transcriptional kinetics, protein competition, cellular concentrations, or three-dimensional genome organization.
6. No cellular R-loop measurement, integration-site measurement, or causal perturbation experiment was performed in this study.

4.6 Future validation

The framework is designed to support future empirical testing. Candidate sequence contexts can be prioritized for direct RNA–DNA hybridization measurements and, separately, for cellular assays that detect R-loops. Additional viral strains and other viral genomes can be analyzed with the same pipeline to determine whether the observed pattern persists beyond HXB2.

5 Conclusion

We established a reproducible sequence-level computational framework designed for application across viral genomes and demonstrated it on 16 predefined regions of HIV-1 HXB2.

Fifteen of 16 regions produced negative model-defined thermodynamic contrasts, with a panel mean of -8.00 kcal/mol and a primary two-sided one-sample test of $t_{15} = -5.87$, $p = 1.54 \times 10^{-5}$.

The direction remained stable after exclusion of the overlapping TAR and R_5LTR regions. The appropriate conclusion is that the selected HIV-1 sequence panel shows a consistent modeled thermodynamic contrast under the specified OligoWalk and DuplexFold procedures.

This is a testable computational prediction, not evidence that cellular R-loops form at these regions or that they control HIV-1 integration or host gene regulation.

Data and code availability

The reference genome is publicly available from NCBI GenBank as K03455.1. The analysis pipeline used fixed coordinates, RNAstructure 6.6 command-line calculations, and Python 3.8.10 statistical processing.

The exact local executable paths are environment-specific and are therefore not encoded as a requirement for reproduction. The manuscript reports the sequence coordinates, calculation definitions, production values, and statistical settings needed to reconstruct the analysis.

Author contributions

Saadat Samadi Dinani: conceptualization, computational analysis, quality control, statistical analysis, and manuscript preparation.

Conflict of interest

The author declares no competing interests.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used a large language model tool in order to assist with L^AT_EX formatting, language editing, and computational code organization. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

- [1] Aguilera A, García-Muse T. R loops: from transcription byproducts to threats to genome stability. *Molecular Cell*. 2012;46(2):115–124.
- [2] Santos-Pereira JM, Aguilera A. R loops: new modulators of genome dynamics and function. *Nature Reviews Genetics*. 2015;16(10):583–597.
- [3] Sollier J, Cimprich KA. Breaking bad: R-loops and genome integrity. *Trends in Cell Biology*. 2015;25(9):514–522.
- [4] Sanz LA, Hartono SR, Lim YW, Steyaert W, Rajpurkar A, Vandeventer PJ, et al. Prevalent, dynamic, and conserved R-loop structures associate with specific epigenomic signatures in mammals. *Molecular Cell*. 2016;63(1):167–178.
- [5] Chedin F. R-loop formation: cause or consequence? *Nature Structural & Molecular Biology*. 2016;23:419–421.
- [6] Crossley MP, Bocek M, Cimprich KA. R-loops as cellular regulators and genomic threats. *Molecular Cell*. 2019;73(3):398–411.
- [7] García-Muse T, Aguilera A. R loops: from physiological to pathological roles. *Cell*. 2019;179(3):604–618.

-
- [8] Petermann E, Lan L, Zou L. Sources, resolution and physiological relevance of R-loops and RNA-DNA hybrids. *Nature Reviews Molecular Cell Biology*. 2022;23(8):521–540.
- [9] Ratner L, Haseltine W, Patarca R, Livak KJ, Starcich B, Josephs SF, et al. Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature*. 1985;313:277–284.
- [10] Korber B, Foley BT, Kuiken C, Pillai SK, Sodroski JG. Numbering positions in HIV relative to HXB2CG. In: Kuiken C, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S, Korber B, editors. *Human Retroviruses and AIDS 1998*. Los Alamos National Laboratory; 1998. p. III-102–III-111.
- [11] Coffin JM, Hughes SH, Varmus HE. *Retroviruses*. Cold Spring Harbor Laboratory Press; 1997.
- [12] Schröder ARW, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*. 2002;110(4):521–529.
- [13] Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Research*. 2007;17(8):1186–1194.
- [14] Lewinski MK, Bushman FD. Retroviral DNA integration—mechanism and consequences. *Advances in Genetics*. 2005;55:147–181.
- [15] Ciuffi A, Mitchell RS, Hoffmann C, Leipzig J, Shinn P, Ecker JR, Bushman FD. Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts. *Molecular Therapy*. 2006;13(2):366–373.
- [16] Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*. 2010;11:129.
- [17] Ali SE, Mittal A, Mathews DH. RNA secondary structure analysis using RNAstructure. *Current Protocols*. 2023;3(7):e846.
- [18] Mathews DH, Burkard ME, Freier SM, Wyatt JR, Turner DH. Predicting oligonucleotide affinity to nucleic acid targets. *RNA*. 1999;5(11):1458–1469.
- [19] Lu ZJ, Mathews DH. OligoWalk: an online siRNA design tool utilizing hybridization thermodynamics. *Nucleic Acids Research*. 2008;36(Web Server issue):W104–W108.
- [20] Mathews DH. Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics*. 2005;21(10):2246–2253.
- [21] Turner DH, Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*. 2010;38(suppl_1):D280–D282.
- [22] Bellaousov S, Reuter JS, Seetin MG, Mathews DH. RNAstructure: web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Research*. 2013;41(W1):W471–W474.
- [23] SantaLucia J Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences USA*. 1998;95(4):1460–1465.

-
- [24] Sugimoto N, Nakano S, Katoh M, Matsumura A, Nakamuta H, Ohmichi T, Yoneyama M, Sasaki M. Thermodynamic parameters to predict the stability of RNA/DNA hybrid duplexes. *Biochemistry*. 1995;34(35):11211–11216.
- [25] Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. Thermodynamics of RNA/RNA heteroduplexes with Watson-Crick base pairs. *Biochemistry*. 1998;37(42):14719–14735.
- [26] Hamperl S, Cimprich KA. The contribution of co-transcriptional RNA:DNA hybrid structures to DNA damage and genome instability. *DNA Repair*. 2014;19:84–90.
- [27] Aguilera A, Gómez-González B. DNA-RNA hybrids: the risks of DNA breakage during transcription. *Nature Structural & Molecular Biology*. 2017;24:165–172.
- [28] Skourti-Stathaki K, Proudfoot NJ, Gromak N. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites. *Molecular Cell*. 2011;42(6):794–805.
- [29] Hatchi E, Skourti-Stathaki K, Ventz S, et al. BRCA1 recruitment to transcriptional pause sites is required for R-loop-driven DNA damage repair. *Molecular Cell*. 2015;57(4):636–647.
- [30] Bhatia V, Barroso SI, García-Rubio ML, Tumini E, Herrera-Moyano E, Aguilera A. BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. *Nature*. 2014;511(7509):362–365.
- [31] Park J, et al. Human immunodeficiency virus-1 induces host genomic R-loops and preferentially integrates its genome near the R-loop regions. *eLife*. 2024;13:RP97348.
- [32] National Center for Biotechnology Information. Human immunodeficiency virus type 1 (HXB2), complete genome; HIV1/HTLV-III/LAV reference genome. GenBank accession K03455.1.