Biochemistry

pubs.acs.org/biochemistry Article

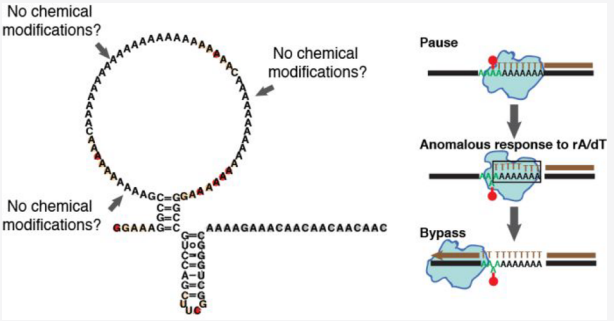
Anomalous Reverse Transcription through Chemical Modifications in Polyadenosine Stretches

Wipapat Kladwang, Ved V. Topkar, Bei Liu, Ramya Rangan, Tracy L. Hodges, Sarah C. Keane, Hashim al-Hashimi, and Rhiju Das*

Cite This: https://dx.doi.org/10.1021/acs.biochem.0c00020 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Thermostable reverse transcriptases are workhorse enzymes underlying nearly all modern techniques for RNA structure mapping and for the transcriptome-wide discovery of RNA chemical modifications. Despite their wide use, these enzymes' behaviors at chemical modified nucleotides remain poorly understood. Wellington-Oguri et al. recently reported an apparent loss of chemical modification within putatively unstructured polyadenosine stretches modified by dimethyl sulfate or 2′ hydroxyl acylation, as probed by reverse transcription. Here, reanalysis of these and other publicly available data, capillary electrophoresis experiments on chemically modified RNAs, and nuclear magnetic resonance spectroscopy on $(A)_{12}$ and variants show that this effect is unlikely to arise from an unusual structure of polyadenosine. Instead, tests of different reverse transcriptases on chemically modified RNAs and molecules synthesized with single 1-methyladenosines implicate a previously uncharacterized reverse transcriptase behavior: near-quantitative bypass through chemical modifications within polyadenosine stretches. All tested natural and engineered reverse transcriptases (MMLV; SuperScript II, III, and IV; TGIRT-III; and MarathonRT) exhibit this anomalous bypass behavior. Accurate DMS-guided structure modeling of the polyadenylated HIV-1 3′ untranslated region requires taking into account this anomaly. Our results suggest that poly(rA-dT) hybrid duplexes can trigger an unexpectedly effective reverse transcriptase bypass and that chemical modifications in mRNA poly(A) tails may be generally undercounted.



RNA molecules play extensive roles in gene regulation and expression in all forms of life.[1] Recent years have seen an explosion of research into the chemical modification of noncoding RNA, mRNA, and synthetic RNA molecules, although the positions and modification rates of bases like m^1A (1-methyladenosine) in human cells remain under intense debate.[2−9] In parallel to these "epitranscriptomic" studies, researchers interested in how RNA molecules fold into elaborate structures measure the reactivity of RNA nucleotides to chemical modification reagents to probe their structural accessibility.[10] To read out the chemical modifications, most modern technologies take advantage of reverse transcriptase (RTase) enzymes, either terminating at modified nucleotides or reading through these positions and leaving mutations, insertions, or deletions in the complementary DNA (cDNA) that correspond to the site of RNA modification. Sequencing of the resulting cDNAs then affords single-nucleotide-resolution measurements of chemical modifications of the original RNA samples.[2,7,11,12]

RTase-based mapping has begun to yield large databases of experimental results, such as the RNA Mapping Database (RMDB) of structure mapping profiles, which catalogs over 50 000 sequences.[13] A surprising finding from these data has been a striking chemical protection signature for long stretches of adenosines (A's).[14] On the one hand, the last (3′-most) six adenosines in such long poly(A) stretches give clear signals for modification by chemical reagents. On the other hand, adenosines 5′ to these last adenosines show no evidence of chemical modification. The observation of this apparent protection pattern arises not only in experiments with dimethyl sulfate (DMS), which methylates chemically accessible N1 positions of adenosines to m^1A, but also in experiments with reagents used for selective 2′-hydroxyl acylation with primer extension (SHAPE), which marks flexible nucleotides (Figure 1a).[15] Substitution of either C, G, or U at any A position results in the restoration of chemical modification signals for 6 nucleotides upstream (5′) but not downstream (3′) of the substitution (Figure 1b).

These observations have potentially significant implications for current efforts to understand RNA structures and chemical

Received: January 7, 2020
Revised: May 11, 2020
Published: May 14, 2020

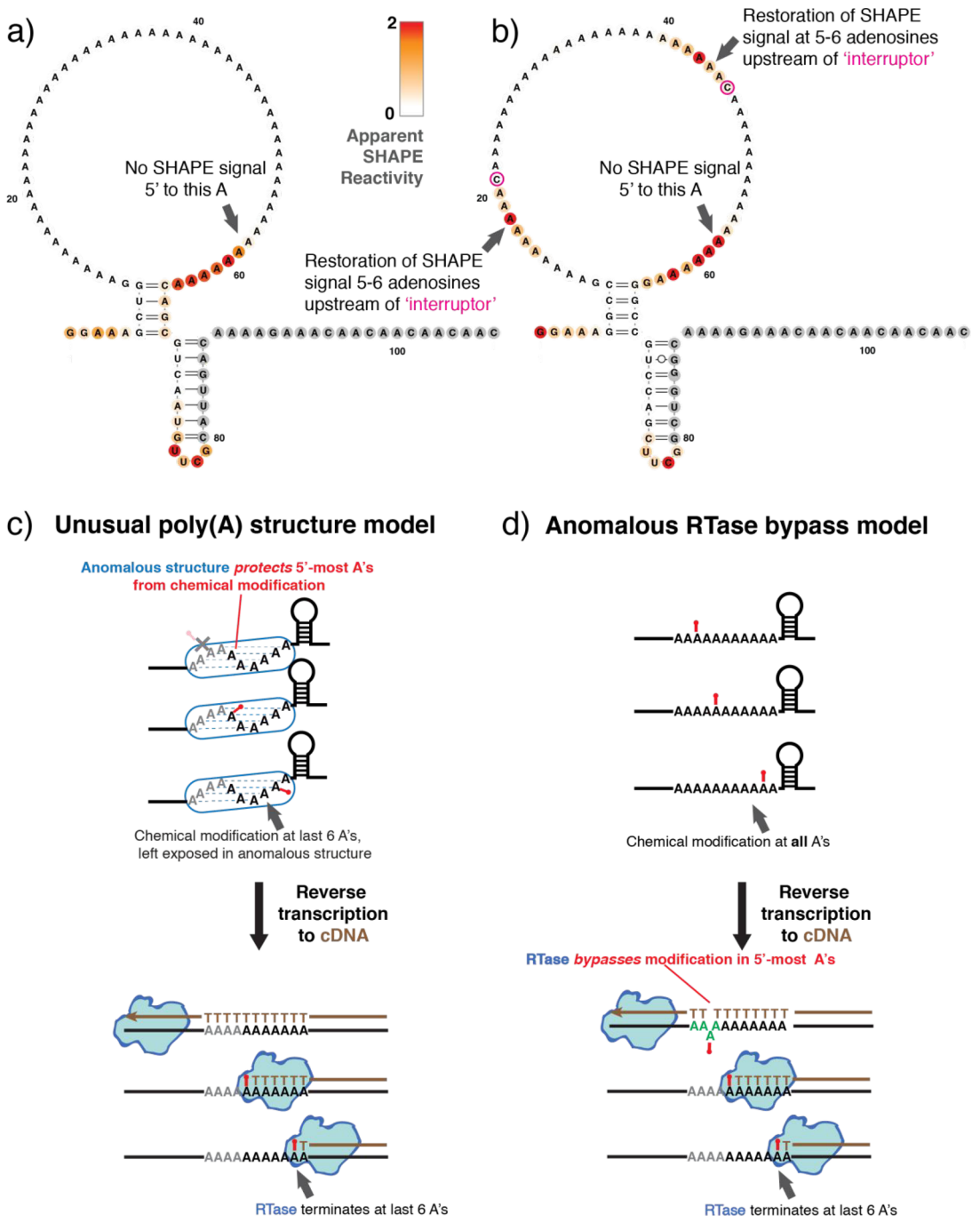


Figure 1. Anomalous signal in poly(A) SHAPE mapping profiles. (a) The SHAPE signal is diminished upstream (5′) of 6 modified adenosines. (b) Installation of cytidine interrupters (magenta) restores the SHAPE signal for the 6 adenosines upstream (5′) to the interrupter. Data report on the termination of SuperScript III reverse transcription at RNA modifications induced by 1M7 (1-methyl-7-nitroisatoic anhydride) and correspond to RNA Mapping Database entry ETERNA_R69_0000, sequence numbers 851 and 852; data normalized so that GAGUA pentaloops in coloaded reference constructs give mean SHAPE reactivities of unity.[25] (c, d) These effects might be accounted for by the presence of an unusual poly(A) structure that protects the 5′-most adenosines from chemical modification (c) or anomalous behavior of the reverse transcriptase in which the enzyme bypasses chemically modified adenosines after polymerizing through 6 modified adenosines (d).

modifications. Long stretches of adenosines appear throughout bacterial and eukaryotic mRNA, especially at their termini,[16,17] and occur in numerous viral sequences, including the end of the HIV genome.[18] Studies that seek to engineer unpaired loops into RNAs typically also use poly(A).[14,19,20] Nevertheless, understanding implications of the anomalous signal requires knowing the mechanistic origin of the poly(A) signature, especially if it is due to a special RNA structure in poly(A) stretches, as was previously hypothesized;[14] or to anomalies in how RTases respond to poly(A),[7,21] analogous to effects seen in poly(A) processing by other molecular machines like the ribosome,[22] RNA polymerase,[23] or the RNase H domains of HIV RTases.[24] Here, we use alternative RNA substrates, capillary electrophoresis, nuclear magnetic resonance spectroscopy, and high-throughput sequencing to understand this mechanism. Our data disfavor the presence of

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

an unusual structure in poly(A) RNA as an explanation of the anomalous signal. Instead, the results implicate an unexpected ability of RTases to bypass chemically modified adenosine after they have already polymerized through stretches of 4−6 adenosines, potentially triggered by an anomalous structure of poly(rA-dT) hybrids in the RTase active site.

METHODS

Enzymatic Nucleic Acid Preparation and Purification. RNA was transcribed using DNA templates containing a T7 RNA Polymerase promoter sequence at their 5′ ends and a 20 base-pair Tail2 sequence at their 3′ ends (Supporting Information). The RNA sequence consisted of the sequence of interest flanked by reference hairpins on each side, serving as internal structural controls. The DNA template was assembled through DNA primer assembly using primers designed using the Primerize tool (https://primerize.stanford.edu/).[26] Designed primers were ordered in plate format from Integrated DNA Technologies (IDT) and assembled via PCR assembly with Phusion DNA polymerase as described on the Primerize PCR Assembly Protocol (https://primerize.stanford.edu/protocol/#PCR). Assembly products were verified for size via agarose gel electrophoresis and subsequently purified using Agencourt RNAClean XP beads. Purified DNA was quantified via NanoDrop (Thermo Scientific), and 8 pmol of purified DNA was then used for *in vitro* transcription with T7 RNA polymerase (New England Biolabs Inc.). The resulting RNA was purified with Agencourt RNAClean XP beads supplemented with an additional 12% of PEG-8000 (3 volumes of 40% PEG-8000 was added to 7 volumes of Agencourt RNAClean XP beads) and quantified via NanoDrop. The sequences of the transcribed Triangle of Doom (TOD) RNA molecules, including reference hairpins, and of transcribed stem-extended TOD-S1 and TOD-S7 RNA molecules, are provided in Supporting Information.

RNA Modification. RNA (1.2 pmol) was denatured in 50 mM Na-HEPES, pH 8.0, at 90 °C for 3 min and cooled at room temperature for 10 min. The RNA was then folded with the addition of $MgCl_2$ to a final concentration of 10 mM in 15 μL and incubated at 50 °C for 30 min, then left at room temperature for 10 min. For chemical modification of folded RNA, fresh working stocks of dimethyl sulfate (DMS, Sigma-Aldrich) and 1-methyl-7-nitroisatoic anhydride (1M7) were prepared. For DMS, 1 μL of DMS was mixed with 99 μL of 100% EtOH. The 100 μL solution was then added to 100 μL of RNase-free H_2O for a final volume of 200 μL. For 1M7, 4.24 mg of 1M7 was dissolved in 1 mL of anhydrous DMSO. In the modification step, 5 μL of the modifying agent (either DMS or 1M7) working stock was added to 15 μL of folded RNA and incubated at room temperature for 15 min. Modification reactions were quenched with 5 μL of quenching solution (β-mercaptoethanol or 0.5 M Na-MES pH 6.0 for DMS and 1M7, respectively). Then, 5 μL of 5 M NaCl, 1.5 μL of oligo-dT Poly(A)Purist MAG beads (Ambion), and 0.065 pmol of 5′ Fluorescein (FAM)-labeled A20-Tail2 primer were added (see Supporting Information), and the solution was mixed and incubated for 15 min. The magnetic beads were then pulled down by placing the mixture on a 96-post magnetic stand, washed twice with 100 μL of 70% EtOH, and air-dried for 10 min before being resuspended in 2.5 μL of RNase-free H_2O.

cDNA Synthesis. The 2.5 μL resuspension of purified, magnetic beads carrying chemically modified RNA was mixed with 2.5 μL of reverse transcription mix with SuperScript III (Thermo Fisher) or an alternate reverse transcriptase as otherwise specified (Table S1). The reaction was incubated at 48 °C for 60 min. The RNA was then degraded by adding 5 μL of 0.4 M NaOH and incubating the mixture at 90 °C for 3 min. The degradation reaction was placed on ice and quickly quenched by the addition of 2 μL of an acid quench solution (1.4 M NaCl, 0.6 M HCl, and 1.3 M NaOAc).

Measurement of Chemical Reactivity with Capillary Electrophoresis. Bead-bound, FAM-labeled cDNA was purified by magnetic bead separation, washed twice with 100 μL of 70% EtOH, and air-dried for 10 min. To elute the bound cDNA, the magnetic beads were resuspended in 10.0625 μL of ROX/Hi-Di (0.0625 μL of ROX 350 ladder, Applied Biosystems, in 10 μL of Hi-Di formamide, Applied Biosystems) and incubated at room temperature for 20 min. The resulting eluate was loaded onto capillary electrophoresis sequencers (ABI-3100 or ABI-3730) either on a local machine or through capillary electrophoresis services rendered by ELIM Biopharmaceuticals (Hayward, CA). Data were analyzed, background-subtracted, and normalized with HiTRACE.[25,27]

Varying Chemical Mapping Experimental Conditions. For both the TOD-S1 and S7 RNA constructs, the chemical mapping experiments described above were carried out under a variety of conditions. The pH of the RNA denaturation and folding steps was varied from 5.0 through 10.0 by using 50 mM of one of the following folding buffers: Na-MES pH 5.0, Na-MES pH 6.0, Tris-HCl pH 7.0, Na-HEPES pH 8.0, Na-CHES pH 9.0, and Na-CHES pH 10.0. For 1M7, final working modifier concentrations of 0.1, 0.52, 0.24, and 1.06 mg/mL in DMSO were tested. For DMS, final modification reaction concentrations of 0.25%, 0.125%, 0.06%, and 0.03% were tested. For folding reaction salt concentrations, 0, 0.1, 0.2, 5.0, 10.0, and 100 mM of $MgCl_2$ or 1.0 and 2.0 M of NaCl (with no $MgCl_2$) were tested. Final concentrations of DMSO in each working modifier stock were tested at 0%, 5%, 10%, 25%, and 50% for DMS; 5%, 10%, 25%, 50% for 1M7. The folding and modification steps were tested with and without the presence of 0.6 pmol of the FAM-labeled A20-Tail2 primer. Modification reaction temperatures of 0, 10, 24, 37, 50, 65, 80, and 98 °C were tested both in the presence and absence of $MgCl_2$. Due to Mg^{2+}-catalyzed RNA hydrolysis at high temperatures, RNA reactivity could not be measured at 98 °C. Several reverse transcriptases, including SuperScript II, Superscript II with $MnCl_2$, SuperScript III, SuperScript IV, TGIRT-III, MarathonRT, AMV, and MMLV, were tested in this study. Reaction conditions for each are listed in Table S1.

Sequencing-Based Chemical Mapping Readout of Terminations and Insertions/Deletions/Mismatches (MAP-seq). To map terminations and mutations incorporated by different reverse transcriptases across from chemically modified TOD-S1 and TOD-S7 RNAs as well as the synthetic Syn41 substrate with m^1A, we used the protocol described in ref 28. For SHAPE and DMS modification, 5.0 pmol of RNA in 4 μL of RNase-free H_2O was denatured by incubating at 95 °C for 2 min and then cooling on ice for 1 min. Then 5.0 μL of 2× buffer (100 mM Na-HEPES, pH 8.0, and 20 mM $MgCl_2$) was added, and the RNA was incubated at 50 °C for 30 min to fold. The RNA was modified by adding 1.0 μL of modifier (10.6 mg/mL of 1M7 in DMSO, 30 mg/mL of NMIA, or 1.25% DMS in 8.75% EtOH). For no modification controls, 1.0 μL of RNase-free H_2O was added instead. Reactions were incubated at 24 °C for 15 min and then quenched with 2 μL of 0.5 M Na-MES, pH 6.0, or β-mercaptoethanol for SHAPE or

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

DMS modification, respectively. Modified RNA was then purified by ethanol precipitation. Reverse transcription and ligation of the second Illumina sequencing adapter were carried out as previously described.[28] The RTB barcode assigned to each sample, and the sequences of each RTB barcoded reverse transcription primer are listed in Tables S2 and S3 and (Supporting Information). For the second adapter ligation, P_TruSeqAdapt01_p and P_TruSeqAdapt02_p ligation primers were used for TOD-S1 and TOD-S7 constructs, respectively (see Supporting Information). Sequencing libraries were quantified by qPCR and sequenced on an Illumina MiSeq using a 150-cycle v3 chemistry kit per manufacturer's instructions, with 35 and 121 cycles used for read 1 and read 2, respectively. Data were analyzed for termination events with MAPseeker software[28] and for mismatch/deletion events with ShapeMapper[12] and M2seq.py software (https://github.com/ribokit/M2seq).[29] Alternative analyses of mismatch/deletion/insertion events were carried out with ShapeMapper 2,[30] RNA Framework,[31] and custom Python scripts (https://github.com/DasLab/Anomalous_polyA_RT).

Chemical Synthesis of Modified RNAs. All chemically synthesized 41-mer TOD-S1 RNA molecules were generated at a 40 nmol scale using the ABI-394 DNA/RNA synthesizer (Thermo Fisher Scientific) at the Stanford Protein and Nucleic Acid Facility (PAN). Through the process, synthesis and purification conditions avoided basic conditions that catalyze Dimroth rearrangement of m^1A to m^6A.[32] For wild-type RNA molecules, UltraFast chemistry was used with Bz-A-CE, Ac-C-CE, Ac-G-CE, and U-CE phosphoramidites (Glen Research catalog numbers 10-3003-10, 10-3015-10, 10-3025-10, 10-3030-10, respectively). For RNAs containing m^1A, UltraMild chemistry was used with Pac-A-CE, 1-Me-A-CE, Ac-C-CE, and iPr-Pac-G-CE phosphoramidites (Glen Research catalog numbers: 10-3000-10, 10-3501-95, 10-3015-10, 10-3021-10, 10-3030-10, respectively). (We could not chemically synthesize RNAs to test SHAPE-induced reverse transcriptase bypass, as the exact 2′-hydroxyl adducts formed by SHAPE chemistry are not available as phosphoramidites.) Synthesized RNA was desalted with a Gel-Pak 1.0 desalting column (Glen Research) and concentrated with a SpeedVac (Thermo Fisher Scientific) on low heat. Concentrated RNA was checked for the appropriate molecular weight with the Voyager-DE STR Biospectrometry Workstation (Applied Biosystems) and a MALDI-TOF mass spectrometer (Applied Biosystems). Quality-verified RNA was finally size selected using a denaturing 10% PAGE gel.

ssRNA Synthesis for NMR Spectroscopy. 12-mer RNA oligonucleotides were synthesized using a MerMade 6 Oligo Synthesizer employing 2′-tBDSilyl protected phosphoramidite (Bz-A-CE Phosphoramidite 10-3003-10, Ac-C-CE Phosphoramidite 10-3015-10, Ac-G-CE Phosphoramidite 10-3025-10, U-CE Phosphoramidite 10-3030-10 from ChemGenes) on 1 μmol standard synthesis columns (1000 Å) (BioAutomation). RNA oligonucleotides were synthesized with the option to leave the final 5′-protecting group (4,4′-dimethoxytrityl (DMT)) on. Synthesized oligonucleotides were cleaved from the 1 μmol column using 1 mL of ammonia methylamine (1:1 ratio of 30% ammonium hydroxide and 30% methylamine) followed by 2 h of incubation at room temperature to allow base deprotection. The solution was then air-dried and dissolved in 115 μL of DMSO, 60 μL of TEA, and 75 μL of TEA-3HF, followed by 2.5 h incubation at 65 °C for 2′O-protecting group removal. Samples were then quenched with Glen-Pak RNA quenching buffer and loaded onto Glen-Pak RNA cartridges (Glen Research Corporation) for purification using the online protocol (http://www.glenresearch.com/). Samples were then ethanol precipitated, air-dried, dissolved in water, and then buffer exchanged at least three times using a centrifugal concentrator (EMD Millipore) into the desired buffer.

NMR Experiments. All NMR experiments were collected on a 600 MHz or a 700 MHz Bruker NMR spectrometer equipped with an HCN cryogenic probe. Data were processed and analyzed using NMRpipe[33] and SPARKY,[34] respectively. Complete assignment experiments were obtained using 2D [^{13}C, ^1H] Heteronuclear Single Quantum Coherence (HSQC), 2D [^1H, ^1H] WATERGATE Nuclear Overhauser Effect Spectroscopy (NOESY, mixing time of 150 ms) experiments.

DMS Profiling and Mutate-and-Map-Seq of HIV 3′UTR. For mutate-and-map-seq (and DMS-MaP-seq), the protocol in ref 29 was followed. RNA constructs for HIV 3′-UTR of reference genome HIV-1 NL4−3 were prepared without and without A_{20} inserted after the polyadenylation site (see Supporting Information). DNA templates (including prepended T7 promoter) were synthesized by PCR assembly based on primers designed by Primerize.[26] PCR assembly was carried out using Phusion DNA polymerase as above, using annealing temperature of 64 °C for 35 cycles. DNA templates for mutate-and-map-seq were prepared with additional mutations by amplifying this DNA under error-prone PCR conditions: 10 mM Tris-HCl, pH 8.3, 50 mM KCl, 4 mM $MgCl_2$, 0.5 mM $MnCl_2$, 1 mM dCTP, 1 mM dTTP, 0.2 mM dATP, 0.2 mM dGTP, 2 μM for forward and reverse primer, template 2 ng, Taq-DNA polymerase 5 units. PCR conditions were as follows: 94 °C, 1 min; then 24 cycles of 94 °C for 30 s, 64 °C for 1 min, 72 °C for 3 min; and 72 °C for 10 min for 1 cycle. DNA templates were purified by Qiagen PCR spin purification columns.

DNA templates were transcribed with T7 RNA polymerase (NEB) at 37 °C for 6 h, and RNA was purified by Agencourt Ampure XP beads supplemented with PEG-8000 (see above). In preparation for DMS chemical mapping, 12.5 pmol of RNA was heated at 95 °C for 2 min, cooled on ice for 1 min and then incubated in 300 mM Na-cacodylate, pH 7.0, 10 mM $MgCl_2$ at 37 °C for 30 min at a final volume of 25 μL. Modification conditions used were 0.17 M of dimethyl sulfate at 37 °C for 6 min. RNase-free water was added instead as a no modification control. DMS reactions were quenched with 25 μL of β-mercaptoethanol. RNA was purified by ethanol precipitation in 10% v/v of 3 M sodium acetate. Purified RNA was resuspended in 7 μL of RNase-free water. Then, 4.6 μL of the RNA sample was used to generate cDNA with SuperScript II reverse transcriptase with Mn^{2+} as the divalent ion to promote the incorporation of mismatches or deletions across bypassed methylations. Reverse transcription reactions included 0.02 μM primer in a final volume of 12 μL, buffer of 25 mM Tris-HCl, pH 8.3, 75 mM KCl, 6 mM $MnCl_2$, and 5 mM DTT and incubation at 42 °C for 3 h. Barcoded primers are listed in Table S4 and Supporting Information. Reverse transcription reactions were stopped with 5 μL of 0.4 M sodium hydroxide at 90 °C for 3 min. Reactions were neutralized with acid quench (2 volumes 5 M NaCl, 2 volumes 2 M HCl, and 3 volumes of 3 M Na−acetate).

cDNA was purified with Agencourt Ampure beads supplemented with PEG-8000 (see above), and purified

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

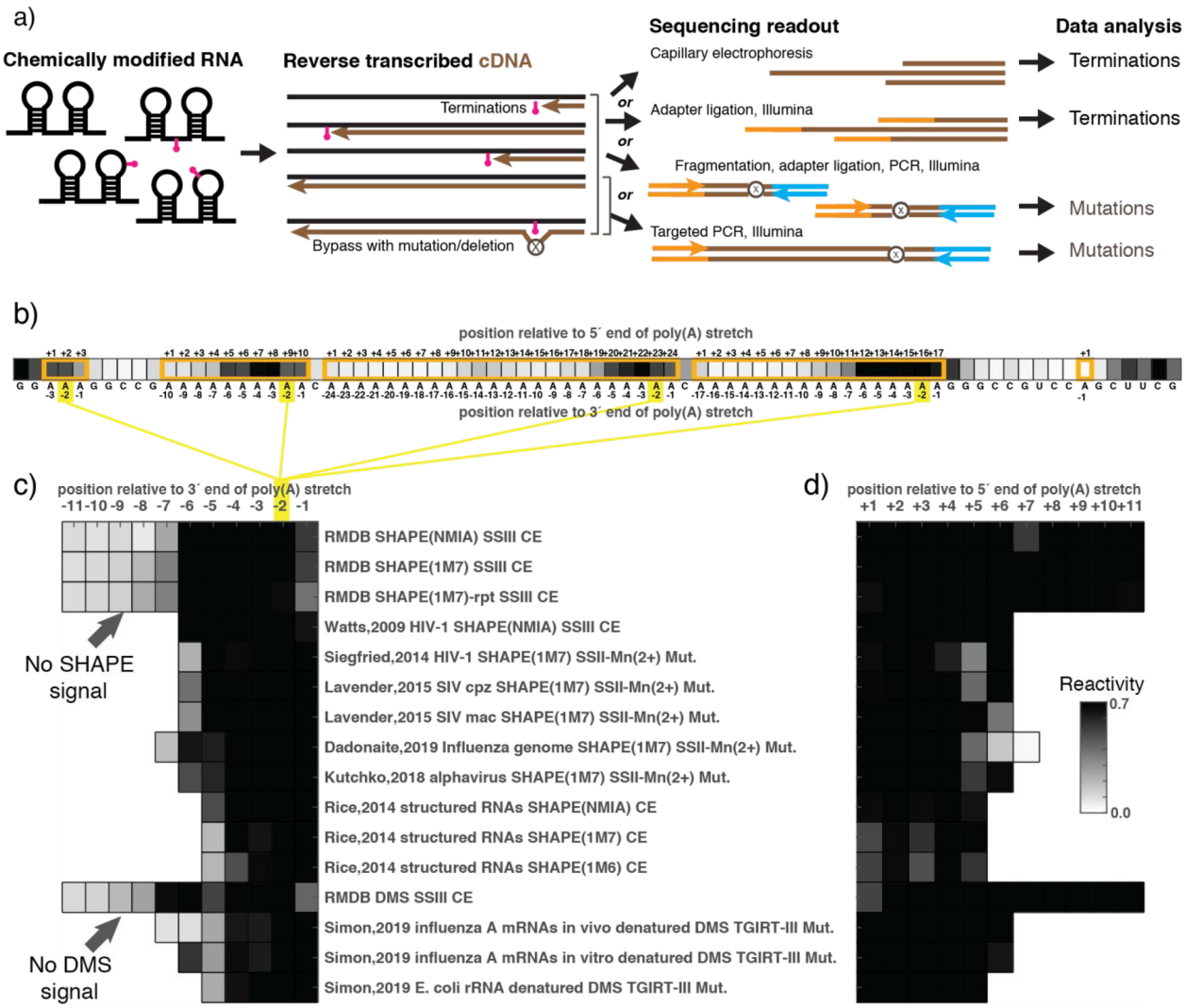


Figure 2. Survey of publicly available chemical mapping data support poly(A) anomaly. (a) Protocols for mapping RNA chemical modifications involve several steps, each with multiple possible variations. (b) Schematic for how adenosines within the single sequence shown in Figure 1b are indexed before averaging DMS reactivity for positions based on positions within each polyadenosine stretch. (c, d) Reactivity data averaged according to the position relative to the 3′ end (c) or 5′ end (d) for each instance of a polyadenosine stretch. Reactivities within each data set were normalized so that the mean reactivity is unity at positions with maximal mean reactivity (typically −1 or +1 positions). References: RMDB (ETERNA_R69_0001);[13,14] Watts, 2009;[38] Siegfried, 2014;[12] Lavender, 2015;[39] Dadonaite, 2019;[40] Kutchko, 2018;[41] Rice, 2014;[37] Simon, 2019.[44] RMDB = RNA Mapping Database, SHAPE = selective 2′-hydroxyl acylation and primer extension, NMIA = *N*-methyl isatoic anhydride, 1M7 = 1-methyl-7-nitroisatoic anhydride, DMS = dimethyl sulfate; SS = SuperScript (reverse transcriptase), SSII-Mn(2+) = SuperScript II reverse transcriptase carried out with Mn^{2+} to promote mutational bypass, TGIRT = thermostable group II intron reverse transcriptase, CE = capillary electrophoresis (reads out reverse transcription termination), Term. = reverse transcription termination read out by Illumina sequencing, Mut. = reverse transcription incorporation of mismatches or deletions read out by Illumina sequencing.

cDNA was resuspended in 12.5 μL of RNase-free water. Then, 2.5 μL of cDNA was used for PCR with Illumina adapters (see Supporting Information). To amplify the full-length cDNA, we found it necessary to perform emulsion PCR. For these reactions, we prepared an oil phase system composed of ABIL EM90 (Evonika), 80 μL, Triton X-100, 1.0 μL, and mineral oil, 1919 μL. Reactions combined 300 μL of this oil phase with 50 μL of the PCR mixture, composed of Phusion DNA polymerase, 1× Phusion buffer, 10 mM dNTPs, and 2 μM of each forward and reverse primer. We performed PCR at 98 °C, 30 s; then 22 cycles of 98 °C for 10 s, 64 °C for 30 s, 72 °C for 30 s, and 72 °C for 10 min for 1 cycle. The resulting dsDNA was purified by gel extraction and purification in Qiagen Qiaquick purification spin columns. The dsDNA was quantified with a Qubit instrument, with the HS dsDNA kit. A Miseq 600-cycle kit was used for sequencing, and data were analyzed with the M2seq py pipeline described in ref 29. Structure prediction was carried out with the *Fold* executable in RNAstructure 6.1, which implements DMS-guided modeling with the pseudoenergy framework described in ref 35.

Data Accessibility. Data for reverse transcribed cDNAs are deposited at the RNA Mapping Database[11] (https://rmdb.stanford.edu) with the following accession IDs.

TOD-S1 RNA, 1M7 (SHAPE) across different solution conditions: TODS1_1M7_0001.

TOD-S1 RNA, DMS across different solution conditions: TODS1_DMS_0001.

TOD-S7 RNA, 1M7 (SHAPE) across different solution conditions: TODS7_1M7_0001.

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

TOD-S7 RNA, DMS across different solution conditions: TODS7_DMS_0001.

TOD-S1 RNA, 1M7 (SHAPE), vary RTases: TODS1_1M7_0002.

TOD-S7 RNA, 1M7 (SHAPE), vary RTases: TODS7_1M7_0002.

TOD-S7 RNA; 1M7, NMIA, DMS; vary RTases; terminations read out by Illumina sequencing: TODS7_TRM_0001.

TOD-S7 RNA; 1M7, NMIA, DMS; vary RTases; mismatches/deletions read out by Illumina sequencing: TODS7_MUT_0001.

TOD-S1 and TOD-S7 RNAs, sequences to extend stems into polyA; 1M7: TODEX_1M7_0000.

TOD-S1 and TOD-S7 RNAs, sequences to extend stems into polyA; DMS: TODEX_DMS_0000.

Syn41 RNA, chemically synthesized with m^1A at different positions, no further modification, varying reverse transcriptases: SYN41_M1A_0001.

HIV 3′ UTR, no A_{20} tail, DMS: HIV3PR_DMS_0001.

HIV 3′ UTR, no A_{20} tail, no modifier control: HIV3PR_NMD_0001.

HIV 3′ UTR, no A_{20} tail, DMS, transcribed off error-prone PCR DNA (mutate-and-map-seq): HIV3PR_DMS_0002.

HIV 3′ UTR, no A_{20} tail, no modifier control, transcribed off error-prone PCR DNA: HIV3PR_NMD_0002.

HIV 3′ UTR, with A_{20} tail, DMS: HIV3PR_DMS_0003.

HIV 3′ UTR, with A_{20} tail no modifier control: HIV3PR_NMD_0003.

HIV 3′ UTR, with A_{20} tail, DMS, transcribed off error-prone PCR DNA (mutate-and-map-seq): HIV3PR_DMS_0004.

HIV 3′ UTR, with A_{20} tail, no modifier control, transcribed off error-prone PCR DNA: HIV3PR_NMD_0004.

Illumina sequencing data are made available at the Gene Expression Omnibus (GEO), Biosample, and Sequencing Read Archive (SRA) databases under overall GEO Series accession ID GSE149061, as the following three sets of the sample (GSM), Biosample (SAMN), and SRA (SRX,SRR) accession IDs.

HIV 3′ UTR, DMS-MaP-seq, M2-seq: SRR11583837, GSM4489497, SAMN14655869, SRX8151467.

TOD-S1 and TOD-S7 RNAs, MAP-seq: GSM4489498, SAMN14655871, SRX8151468, SRR11583838, SRR11583839.

Syn41 RNA, chemically synthesized with m^1A at different positions, no further modification, varying reverse transcriptases: GSM4489499, SAMN14655870, SRX8151469, SRR11583840, SRR11583841, SRR11583842.

MATLAB and Python analysis scripts are available through a Stanford Digital Repository archive at https://purl.stanford.edu/xm553fz5401 and Github repository https://github.com/DasLab/Anomalous_polyA_RT.

RESULTS

Generality of Anomalous Poly(A) Signal Across Prior Data Sets. The suppression of apparent chemical reactivity in poly(A) was initially reported in ref 14 based on RMDB-deposited studies measuring SHAPE and DMS modification with the MAP-seq (multiplexed adduct probing read out by sequencing) method.[28] To test that the anomaly was not an artifact of a single study or method, we compared data sets from numerous independent studies that used a variety of RTases and sequencing techniques (Figure 2a). We cataloged each adenosine stretch in these sequences and compared reactivities of adenosines that were at analogous positions relative to the end of each stretch, indexed as −1 (last adenosine in stretch), −2 (second to last adenosine in stretch), etc. As control comparisons, we also compared the mean reactivities of adenosines by their position relative to the beginning of each stretch (+1, +2, ..., positions; Figure 2b). The RMDB MAP-seq data only show the suppression of chemical reactivity for adenosines at least 6 nucleotides 5′ to the end of poly(A) stretches (−7, −8, etc., in Figure 2c; compare to +7, +8, etc., in 2d).

We first investigated SHAPE data sets, for which the most data are currently available. On the one hand, transcriptome-wide SHAPE studies across millions of nucleotide positions[36] remain sparse, giving zero counts at many positions, and do not yet have the necessary signal-to-noise to check for the poly(A) signal. On the other hand, SHAPE studies focusing on specific RNAs, such as viral genomes (HIV and SIV, alphavirus, and influenza genomes; around 10 000 nucleotides each) or collections of structured RNAs,[37] have higher signal-to-noise.[12,38−41] In several of these data sets, we detected suppression of the SHAPE signal at −6 and −5 positions of polyadenosine stretches (labeled "Siegfried,2014", "Lavender,2015", "Dadonaite,2019", and "Rice,2014" in Figure 2c). However, the data were sparse, and control comparisons (Figure 2c) suggested that the apparent protections might be explained by other effects such as the base pairing of long adenosine stretches or biases in sequencing.[42] Furthermore, there were few or no naturally occurring 7- or 8- adenosine stretches across these viral data sets, which would allow direct comparison to the MAP-seq SHAPE data that only show suppressions at −7, −8, etc. (arrow in Figure 2c).

Beyond SHAPE, there is a growing body of publicly available DMS data on natural RNAs,[11,43] although nearly all that we checked were too poor in signal-to-noise to confirm or falsify the poly(A) anomaly. Fortunately, a recent study acquired DMS profiles for influenza A mRNAs and *E. coli* rRNA with excellent signal-to-noise.[44] These data showed suppression of the DMS signal at adenosines positioned at −5 within polyadenosine stretches (labeled "Simon, 2019" in Figure 2c), as did the original RMDB data studied in ref 14 (Figure 2c), and the control comparisons show no analogous suppression at +1, +2, ..., to +6 positions (Figure 2d). Furthermore, this study carried out DMS modification under denaturing conditions. These data, therefore, hinted that the suppression of the DMS signal might not be due to RNA base pairing, which would be disrupted in those experiments (Figure 1c), but to later steps in the readout of the chemical modifications, as schematized in Figure 1d and further tested below. These data also enabled a comparison of the DMS signal across nearly all 5-mer sequences, which confirmed a striking specificity for a drop in DMS signal at position −5 within polyadenosine stretches and not in other purine-rich 5-mers (Figure S1).

In these studies of published data sets, comparing which adenosine positions showed apparent suppression in the different studies was complicated by the use of different RTases across the studies; different conditions of chemical modification including, in some cases, which reagent was used; differences in whether RTase termination or mismatch-introducing bypass was used to infer chemical modification; and differences in whether capillary electrophoresis or Illumina sequencing was used in the studies (Figure 2a). To more

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

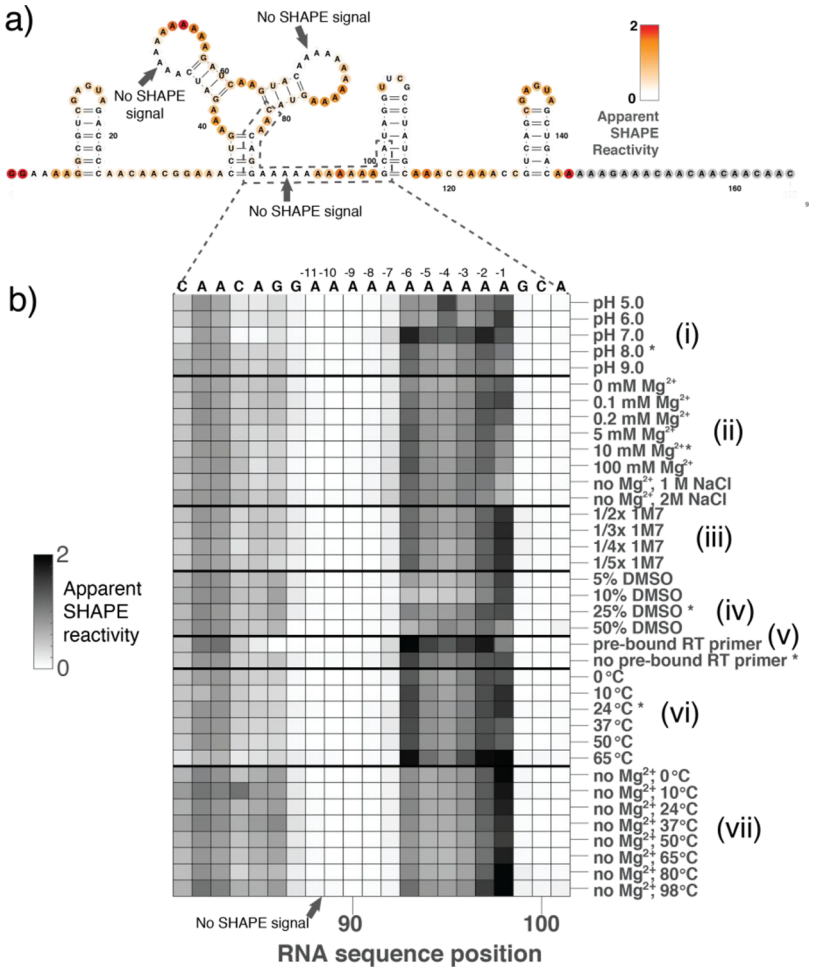


Figure 3. Anomalous SHAPE signals in the TOD-S1 model RNA. (a) Apparent SHAPE reactivity of the TOD-S1 construct, which contains three stretches of 11 adenosines. (b) Anomalous SHAPE profile is robust to a wide variety of changes to solution conditions under which TOD-S1 RNA was chemically modified. Data for the most clearly resolved (3′-most) A11 stretch of the TOD-S1 construct are shown; see Figure S2 for complete profiles. Standard solution conditions for chemical modification were 10 mM $MgCl_2$, 50 mM Na-HEPES, pH 8.0, at an ambient temperature (24 °C), with cosolvent of 25% anhydrous DMSO used to the SHAPE reagent 4.24 mg/mL 1M7. Perturbations to these conditions are labeled on the *y*-axis of panel b, with asterisks (*) marking replicates of the standard conditions. Data shown have been background-subtracted based on control reactions without the SHAPE modification reagent, normalized so that 1.0 corresponds to the average reactivity of GAGUA hairpin loops included at 5′ and 3′ ends of the RNA as internal controls. Anomalous regions with no SHAPE signal highlighted with gray arrows. Data reflect the termination of reverse transcription read out by capillary electrophoresis.

directly confirm the anomalous signal, to dissect its origins, and to characterize the influence of different RTases, protocols, and readouts, we therefore turned to designed model systems.

Anomalous Poly(A) Chemical Mapping Signal Is Recovered with the Simplest Possible Readout. As is apparent from the cross-study analysis of Figure 2, dissecting the mechanism of the anomalous poly(A) chemical mapping signal is complicated by the large number of steps required in current structure mapping methods. The loss of signal at A's could occur at any (or several) of these steps. For example, focusing on the chemical modification step, a single-helix structure of poly(A) with a period of six that leaves its 3′-most turn accessible to chemical modification could explain the observed SHAPE and DMS signals (see Figure 1c and ref 14). Other steps with potential bias include the synthesis of the RNA by *in vitro* transcription with T7 RNA polymerase, which is known to "slip" on repeated sequences;[23] similar slippage or biased termination of reverse transcriptases in the RTase step after chemical modification;[45,46] the ligation of adapters onto the resulting cDNAs, which are known to have sequence biases;[28,43,47] Illumina sequencing-by-synthesis of cDNAs used, which is also known to have sequence biases;[42] or the bioinformatic workup of these sequencing data into inferred chemical modification profiles.[42,48]

We first sought to test if the poly(A) SHAPE and DMS signatures were due to the many possible biases occurring any steps downstream of reverse transcription, by carrying out the simplest method of reading out the cDNA products, capillary electrophoresis with fluorescently labeled primers. This readout detects RTase termination as shorter cDNA products. As a model sequence, we focused on an RNA developed in the Eterna project[14,49] called the Triangle of Doom-Sequence 1 (TOD-S1) (Figure 3). TOD-S1 includes three stretches of 11 A's each: two stretches are designed to be within the loops of hairpins jutting out of a three-way junction, and the third stretch is designed to be part of a single-stranded region near the 3′ end of the construct. Flanking hairpins provide normalization standards (Figure 3a).[25] Unlike the MAP-seq protocol, where RNAs were synthesized in a large pool with other molecules, these RNAs were synthesized individually by

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

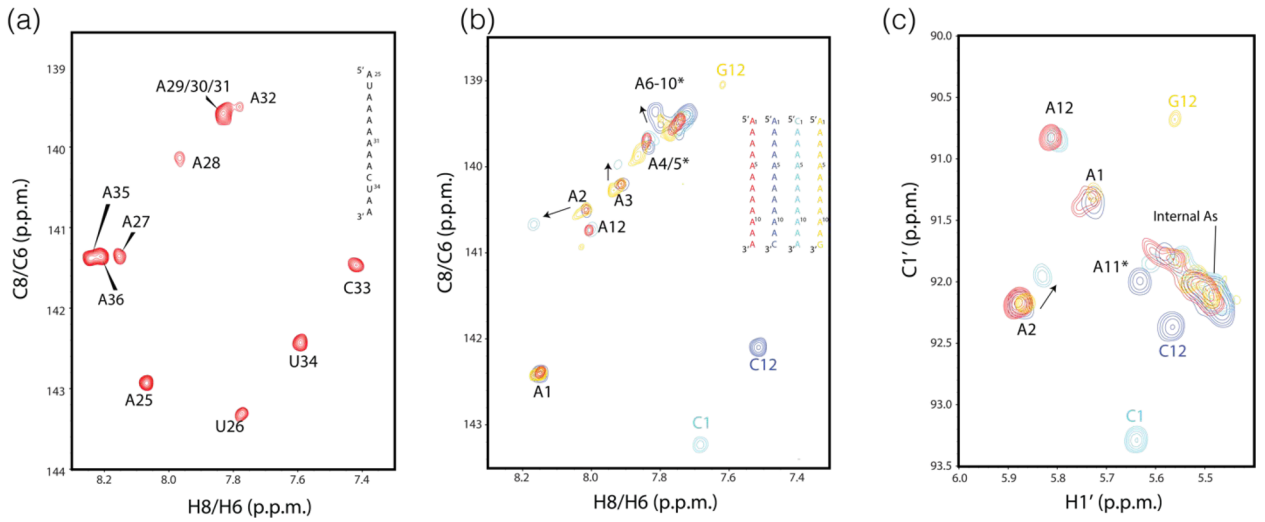


Figure 4. NMR spectra indicate A-form like conformations throughout polyadenosine. (a, b) Chemical shifts of C8 and H8 atoms in [^{13}C,^{1}H] HSQC experiments fall within the range of A-form helices for both a reference sequence previously shown to form a single-stranded stacked configuration[52] (a) and for a 12-adenosine RNA A_{12} (b). In panel b, substitutions of the 5′ adenosine of A_{12} to cytidine (CA_{11}, cyan) or the 3′ adenosine to cytidine ($A_{11}C$, blue) or guanosine ($A_{11}G$, gold) give small perturbations to chemical shifts of immediately neighboring bases (arrows) but no evidence for dramatic perturbations or induction of conformational rearrangements extending six nucleotides. (c) Chemical shifts in the aromatic base region from the same [^{13}C,^{1}H] HSQC experiments also appear in the range of conventional A-form helices. Spectra acquired in Na-HEPES buffer (1 mM RNA in 50 mM Na-HEPES pH 8.0, 10 mM $MgCl_2$) at 25 °C.

T7 RNA polymerase from synthetic DNA templates. After chemical modification and reverse transcription, we directly visualized the cDNA products through capillary electrophoresis (CE), instead of ligating on adapters and carrying out Illumina next-generation sequencing.

Figure 3a shows the CE-based SHAPE data as coloring on the TOD-S1 secondary structure. For each stretch of 11 A's, there was clear evidence for strong SHAPE modification for the last 6 A's but few or no detectable cDNA products corresponding to the modification of the first 5 A's of the same stretch. As above, we have numbered positions in a poly(A) stretch relative to the 3′ end; so here only positions −6 through −1 appeared SHAPE-modified, whereas positions −11 to −7 did not show terminated cDNA products corresponding to SHAPE modification. These patterns corresponded well to analogous measurements in the RMDB MAP-seq experiments.[14] Further CE-based data involving constructs with C mutations in the TOD-S1 loops (called TOD-S7), and using DMS instead of SHAPE, further matched the observations in the prior MAP-seq data (Figure S2). These CE data demonstrated that the anomalous poly(A) signatures observed in MAP-seq and other literature data sets (Figure 2c) were not due to biases in the Illumina sequencing workup of cDNAs since those workup steps are absent in this capillary electrophoresis readout.

Alternative Solution Conditions during Chemical Modification Give the Same Poly(A) Anomaly. The lack of terminated cDNA's corresponding to SHAPE or DMS modification at the middle of the poly(A) could be due to an unusual conformation at those positions that might sterically protect the 2′-hydroxyl or N1 position of the adenosine from chemical modification (Figure 1c). We reasoned that such a poly(A) structure may be destabilized or at least modulated by solution conditions and carried out six sets of experiments to test this prediction (Figure 3b).

At high RNA concentrations and low pH, poly(A) strands are known to form a parallel double helix, which is stabilized by protonation at the adenosine N1 position.[50] We varied the pH of our measurements away from our standard pH of 8.0 from 5.0 to 10.0; however, we saw no change in the poly(A) SHAPE profiles of TOD-S1 (Figure 3b-(i)). Second, RNA secondary and tertiary structures are generally stabilized by the presence of Mg^{2+}, which was present at a concentration of 10 mM $MgCl_2$ in all of the above measurements, and the stability can be modulated by high concentrations of monovalent salt. However, when we repeated our measurements without Mg^{2+} or with lower concentrations of Mg^{2+}, or with 1 or 2 M NaCl (and no Mg^{2+}), we again saw no change in SHAPE profiles of TOD-S1 (Figure 3b-(ii)) Third, we asked if the modification reagent or its cosolvent (anhydrous DMSO) might be stabilizing an alternative poly(A) structure, so we varied the concentration of the SHAPE modifier 1M7, going as low as one-fifth of our standard concentration, but saw similar patterns (Figure 3b-(iii)). Fourth, we directly added the SHAPE cosolvent DMSO to up to 50% concentration, and again saw no change in the poly(A) regions, though we observed perturbed chemical modification profiles outside those regions confirming the denaturing effect of DMSO (Figure 3d-(iv); Figure S2a,b). Fifth, our constructs for MAP-seq and the current CE-based measurements all include an A-rich primer binding site at the 3′ end of the sequence (gray nucleotides, Figures 1a,b and 3a). To test if this region might be interacting with poly(A) sequence, we repeated measurements with the reverse transcription primer present during chemical modification as a competitive inhibitor of a possible interaction; however, we observed the same pattern as before (Figure 3b-(v)). Sixth, we reasoned that if poly(A) forms an unusual structure, we would be able to modulate its stability relative to an unfolded configuration by increasing the temperature. However, in SHAPE measurements ranging from 0 to 98 °C, we observed no changes in the poly(A) regions, either with or without $MgCl_2$ (Figure 3b-(vi) and Figure 3b-(vii)). Data from these six sets of experiments strongly disfavored a model where the observed SHAPE

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

signature at poly(A) might be due to an anomalous RNA structure. Further CE-based experiments with the TOD-S7 variant and using DMS instead of SHAPE gave additional support to these conclusions (Figure S2a,b).

NMR Spectroscopy Further Disfavors the Presence of an Unusual Poly(A) RNA Structure. As a further test of whether poly(A) forms an anomalous structure, we synthesized model RNA molecules and carried out NMR spectroscopy under conditions matching our chemical mapping experiments (50 mM Na-HEPES pH 8.0, 10 mM $MgCl_2$; 25 °C). For an RNA consisting of 12 adenosines, all of the resonances fell in the 2D [^{13}C,^{1}H] HSQC chemical shift region, similar to results for a model RNA shown to adopt a stacked conformation that is in equilibrium with the unfolded state (compare red contours in Figure 4a,b; ref 51). In the SHAPE and DMS chemical mapping experiments described above, the substitution of individual A's to C, G, or U produced large changes in apparent chemical reactivity for stretches of six adenosines 3′ to the substitution (Figure 1). By NMR, however, such substitutions gave only small changes in chemical shifts, and these changes were limited to adenosines immediately neighboring the substitution and not for longer stretches of six adenosines 3′ to the substitution (Figure 4b,c). In addition, 2D NOESY spectra indicated that the RNA's glycosidic bonds all adopt a predominantly anti conformation with relatively weak H1′-H8 NOE cross peaks (Figure S3). Thus, NMR experiments disfavor a noncanonical RNA structure as an explanation for the anomalous chemical mapping signals in polyadenosine, in agreement with the biochemical data above.

Alternative Reverse Transcriptases Give Distinct Poly(A) Signatures. Given the lack of evidence for adapter ligation or Illumina sequencing biases (Figure 3) or an anomalous poly(A) structure (Figures 3 and 4) as an origin of the SHAPE/DMS anomaly, we developed experiments to instead test for unexpected behavior in the intermediate step of reverse transcription (Figure 2a). We considered a model in which the RTase bypasses chemically modified adenosines that are positioned at least six nucleotides ahead of the end of a poly(A) stretch (Figure 1d). In this model, the RTase does not terminate or make a mutation, but instead skips by correctly polymerizing a dT or several dT's (insertion) as it bypasses the chemically modified A; any or all such events would lead to the absence of a termination product detected by capillary electrophoresis and explain the anomalies of Figures 1−3.

One prediction from this RTase bypass model is that different reverse transcriptases could give different cDNA profiles when reverse transcribing the same pools of chemically modified RNAs. We, therefore, tested the Moloney Murine Leukemia Virus (MMLV) reverse transcriptase and its commercially available variants SuperScript II, III, and IV. On the one hand, MMLV and SuperScript II both gave similar patterns to our original standard RTase SuperScript III, giving cDNA termination products at positions −6, −5, to −1 for SHAPE-modified RNA (Figure 5, top). In contrast, SuperScript IV gave cDNA termination products only at two positions −2 and −1, i.e., the last two adenosines of each poly(A) stretch, and efficiently bypassed modifications at more 3′ positions. Interestingly, the behavior of SuperScript IV was nearly identical to the behavior of SuperScript II reverse transcribing in the presence of Mn^{2+}, conditions recommended for mutational bypass.[12] This observation suggests that the MMLV-derived enzymes have an alternative "mode" of bypass that can be triggered either by Mn^{2+} or by mutations installed

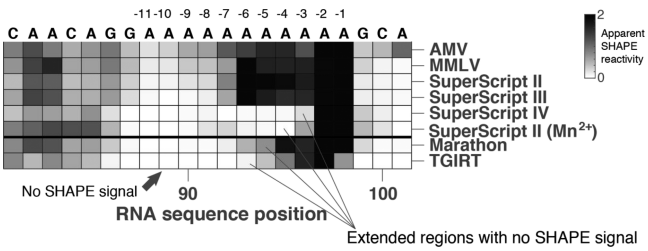


Figure 5. Anomalous SHAPE signals change with reverse transcriptase used to read out chemical modifications. Data for the most clearly resolved (3′-most) A11 stretch of the TOD-S1 construct are shown. AMV = avian myeloblastosis virus. MMLV = Moloney murine leukemia virus. TGIRT = thermostable group II intron reverse transcriptase (variant TGIRT-III). Data reflect the termination of the reverse transcription read out by capillary electrophoresis.

into SuperScript III to create SuperScript IV. Going beyond the MMLV family, we further tested two RTases derived from thermostable group II introns, TGIRT-III,[11,47] and MarathonRT.[53] Both of these enzymes also nearly completely bypassed SHAPE modifications in poly(A) stretches (Figure 5). Additional experiments with the TOD-S7 variant confirmed these conclusions (Figure S2c).

To test readouts different from capillary electrophoresis (Figure 2a) and to test responses to DMS-modified as well as SHAPE-modified RNA, we used adapter ligation followed by Illumina sequencing and discovered analogous differences in termination and in mismatch-free bypass in poly(A) depending on which RTase was used (Figure S4), further implicating the RTase as a likely culprit for the poly(A) anomaly. As a final series of SHAPE-based tests, we designed new constructs that introduced stretches of U's that were distal to the poly(A) stretches of TOD-S1 but base paired to these originally single-stranded poly(A) stretches. The resulting perturbed structures were predicted to give distinct SHAPE patterns in the anomalous structure vs RTase bypass models (Figure S5a−d). As shown in Figure S5e, the actual experimental results strongly favored the RTase bypass model and disfavored the anomalous structure model, in agreement with our previously described analyses (Figures 1−5).

Near-Quantitative Bypass at m^{1}A Modifications Installed during Chemical Synthesis. All of the structure mapping results above from prior studies and from our newer experiments are consistent with near-quantitative RTase bypass within poly(A) stretches. However, these experiments relied on a relatively uncontrolled step: installation of chemical modifications at random positions throughout the RNA molecules via SHAPE or DMS treatment. To remove this randomness, we sought to install chemical modifications specifically at each position of a poly(A) stretch. The RTase bypass model makes predictions for how the RTase should terminate or bypass chemical modifications at each position. For example, DMS modification results in m^{1}A nucleotides. Our DMS results above would therefore predict that SuperScript III RTase should terminate at m^{1}A's installed at the −6, −5, −4, −3, −2 to −1 positions, with particularly strong termination at −3 and −2 and weakest termination at −5 (Figure 2c). The RTase should bypass m^{1}A's installed at any positions 5′ to these last 6 adenosines (Figures 1−3). We tested these predictions on chemically synthesized substrates with m^{1}A at each position in an 11-adenosine stretch (Figure 6). For these experiments, we shifted to a smaller 41-

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

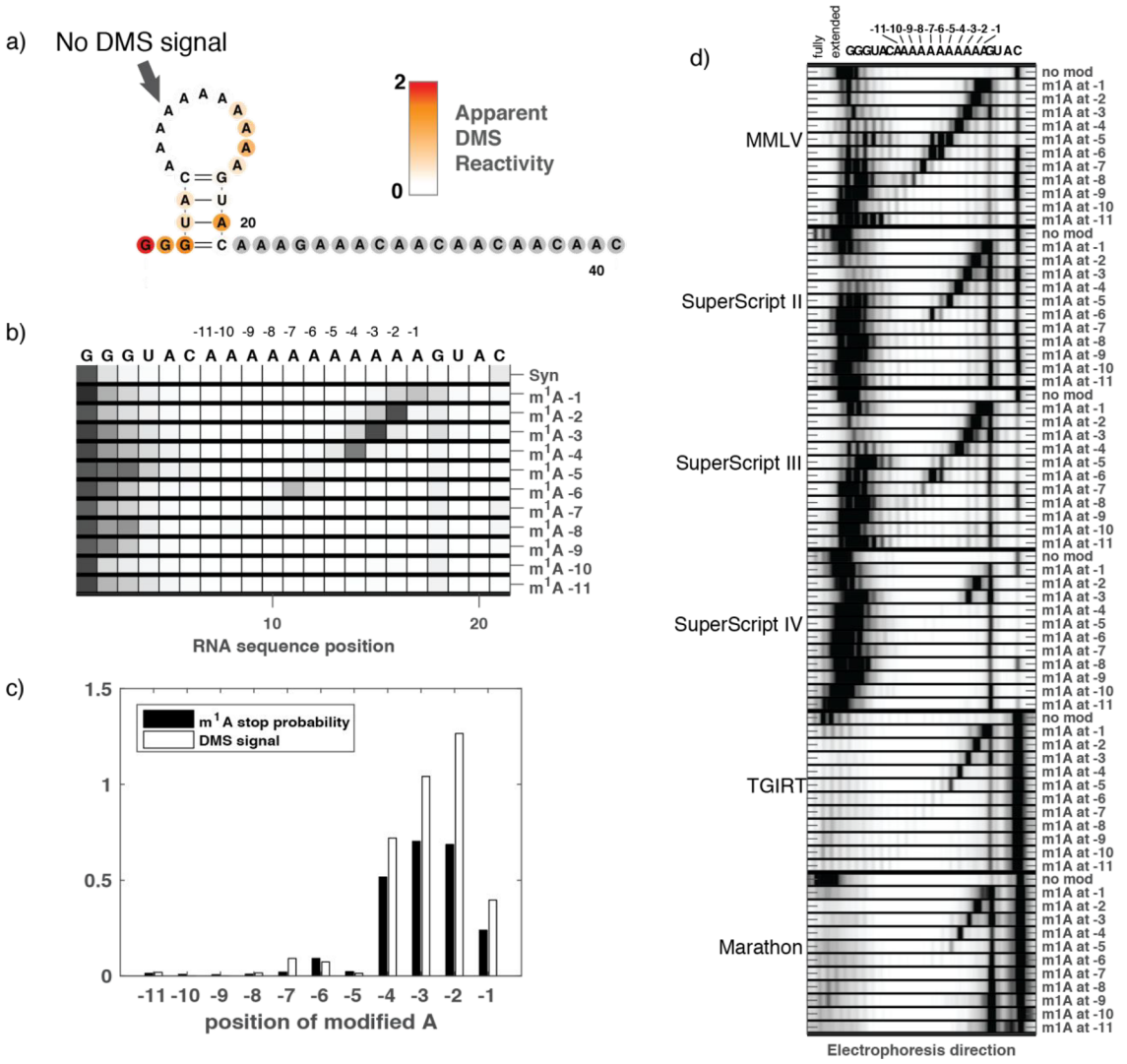


Figure 6. Reverse transcriptase bypass of 1-methyl-adenosines incorporated into a poly(A) stretch during synthesis. (a) DMS mapping data on the Syn41 construct. Data reflect the termination of reverse transcription by SuperScript III reverse transcriptase, read out by capillary electrophoresis. (b) Termination probabilities of SuperScript III reverse transcriptase on constructs that each present a single m^1A installed at the positions specified on the *y*-axis. Note that modifications at −11 through −8 do not appreciably terminate the reverse transcriptase. (c) The measured termination probabilities explain the anomalous signal seen in dimethyl sulfate probing experiments read out by SuperScript III reverse transcriptase, where no signal is apparent at −11 to −8, and the maximum signal appears at −4 to −2. (d) Electropherograms of cDNA extended from Syn41 that is not modified (no mod) or m^1A-modified, as reverse transcribed by different RTases. Sequence assignment of bands is tentative for longer products (toward left of electropherograms), which may contain insertions and deletions. Data reflect the termination of reverse transcription.

nucleotide RNA construct called Syn41 to ensure reasonable synthesis yields and confirmed that chemically synthesized Syn41 gave a poly(A) DMS and SHAPE pattern similar to our previous enzymatically synthesized substrates (Figure 6a).

The reverse transcription results across m^1A-harboring RNAs are shown in Figure 6b. The data agree well with the predictions of the RTase bypass model. SuperScript III terminated nearly quantitatively at m^1A installed at the −3 and −2 positions, with very little full-length cDNA product appearing. The enzyme partially terminated and partially read through m^1A at the other six 3′-most positions of the 11-adenosine stretch. Strikingly, SuperScript III bypassed m^1A installed at −11, −10, −9, −8, and −7 positions with near-quantitative efficiency, resulting in near-full-length cDNAs (Figures 6b–d). This pattern of bypass efficiencies was in excellent agreement with the DMS patterns at poly(A) stretches measured in earlier experiments (Figure 6c). Experiments with other reverse transcriptases on these m^1A substrates also gave bypass efficiencies that accorded with DMS patterns measured previously in other laboratories or by us (compare Figures 2 and 4 to Figure 6d). For example, the group II intron-derived RTases (TGIRT-III and MarathonRT) efficiently reads through m^1A at all positions except −4, −3, −2, and −1 (Figure 6d), similar to their behaviors in literature (labeled “Simon,2019” in Figure 2) and our CE-based DMS experiments (Figure S2c). These comparisons gave strong support for the RTase bypass model, based on prospective experiments.

These experiments also gave new information on how the RTase might bypass chemical modification. Deletions, incorporation, or insertions across from m^1A would give rise to extended cDNA products shorter, the same size, or longer

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

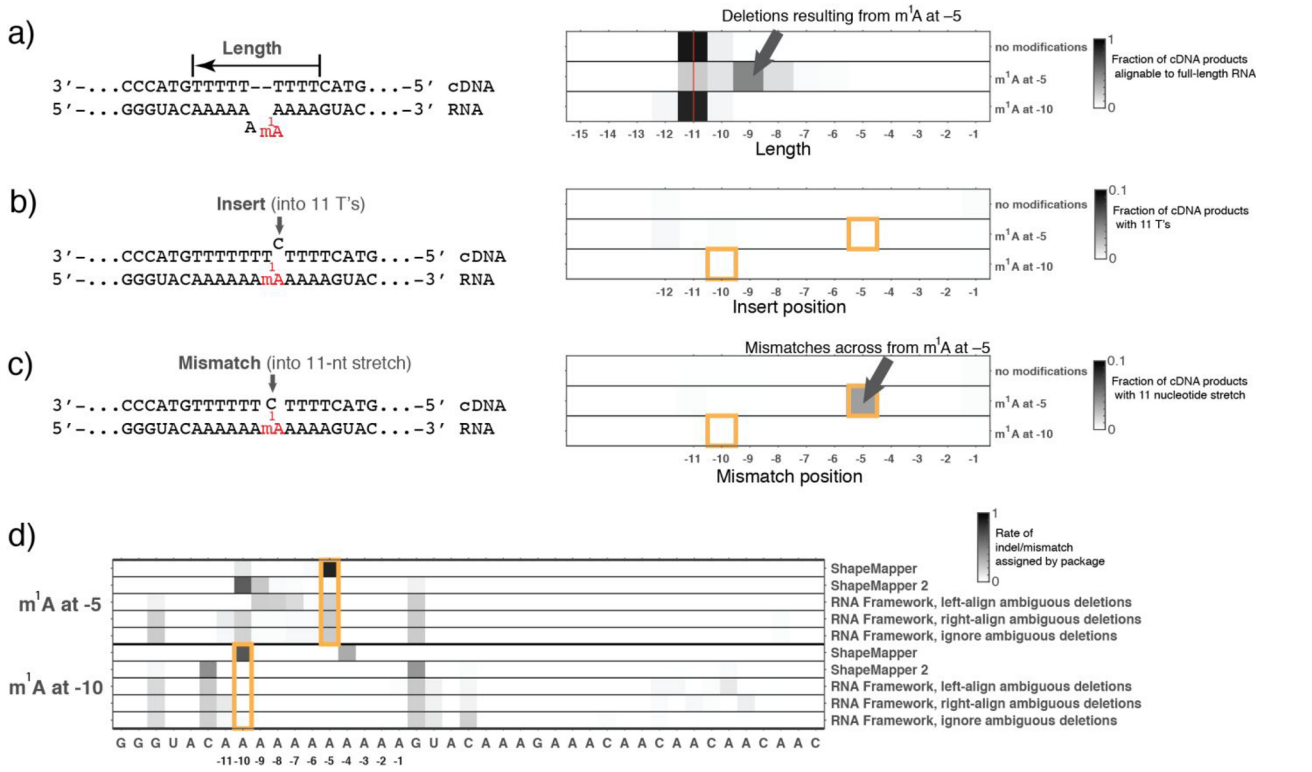


Figure 7. Profiling deletions, insertions, and mismatches introduced by reverse transcription across from m^{1}A in poly(A) stretches. Data are for SuperScript III reverse transcription of the chemical synthesized Syn41 RNA substrate. (a) Lengths of cDNA introduced across from A_{11} stretches, identified based on cDNA matches to immediately flanking sequences. (b) Positions of inserts within cDNA stretches containing exactly 11 T's. (c) Positions of mismatches within cDNA stretches with a length of exactly 11. (d) Output of different analysis packages on the single m^{1}A data sets, using data on the unmodified substrate for background subtraction. Normalization is based on the interquantile method.[35] In panels b–d, gold rectangles highlight locations of mismatches/indels that would correspond to an ideal mutational readout of m^{1}A modification by reverse transcription. Results from other reverse transcriptases are given in Figure S6.

than products reverse transcribed from unmodified RNA. For the MMLV/Superscript family of reverse transcriptases (Figure 6d, left), we observed products that were 1–5 nucleotides shorter in m^{1}A-containing substrates compared to unmodified substrates, suggesting that these enzymes bypass m^{1}A with deletions. Interestingly, the distribution of cDNA lengths depended on the position of the chemical modification in the poly(A) stretch, with m^{1}A at −5 giving not only shorter bypass products but also a spread in termination products (see, in particular, MMLV, Figure 6d) that extend 1 or 2 nts beyond the m^{1}A rather than cleanly terminating at the modification site. Intriguingly, the group II intron-derived RTases (TGIRT-III and MarathonRT) exhibited a near-quantitative bypass of m^{1}A installed at most positions, but the extended products appeared as fainter bands spread across a wider and longer distribution of lengths than the MMLV/Superscript RTases (Figure 6b). The different distributions of bypass products indicate that the molecular details of bypasses are different between enzymes.

Complex Reverse Transcriptase Mutation Profiles across from m^{1}A Modifications. In the experiments above, chemically modified adenosines became invisible to the capillary electrophoresis readout due to the RTase bypass. However, these modifications might still be detectable through mutational events (insertions, deletions, mismatches) recorded in cDNAs and read out by Illumina sequencing of products.[11,12] We carried out such sequencing on RTase products from Syn41 substrates synthesized without modifications, with m^{1}A installed at position −5 and with m^{1}A installed at position −10. We analyzed these data with custom scripts that filtered cDNA's for exact matches of the complete sequences before and after the A_{11} stretch of Syn-41 and searched for deletions, insertions, or mismatches across from A_{11} stretch. Data for SuperScript III are shown in Figure 7, with results for other reverse transcriptases in Figure S6.

With respect to deletions, we primarily detected stretches with lengths of 11 in the unmodified control but a shift to shorter length cDNAs (deletions) for the m^{1}A-5 substrate (Figure 7a). This shift in lengths was expected from our CE experiments (Figure 6b) and was consistent with the bypass scheme of Figure 1d. For the m^{1}A-10 substrate, however, the RTases did not produce shortened products, indicating that bypass occurs without deletions (Figure 7a). This result was also seen in our CE experiments (Figure 6d).

We carried out further filters to cleanly evaluate insertion and mismatch misincorporation in response to the m^{1}A nucleotides. To assess insertions, we filtered the cDNAs for reads that contained exactly 11 T's but potentially other additional nucleotides. We observed a negligible rate of insertions across from m^{1}A in the modified substrates for all RTases (SuperScript III, Figure 7b), except for TGIRT-III in the m^{1}A-5 substrate (Figure S6). We detected no insertions across from m^{1}A-10 for any of the RTases (Figure S6). Finally, we evaluated the incorporation of mismatches across from the m^{1}A nucleotides. To simplify the analysis, we filtered for cDNAs whose segments across from the polyadenosine stretch were exactly 11 nucleotides in length. For all of the RTases tested, we detected non-T nucleotides (mismatches) across

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

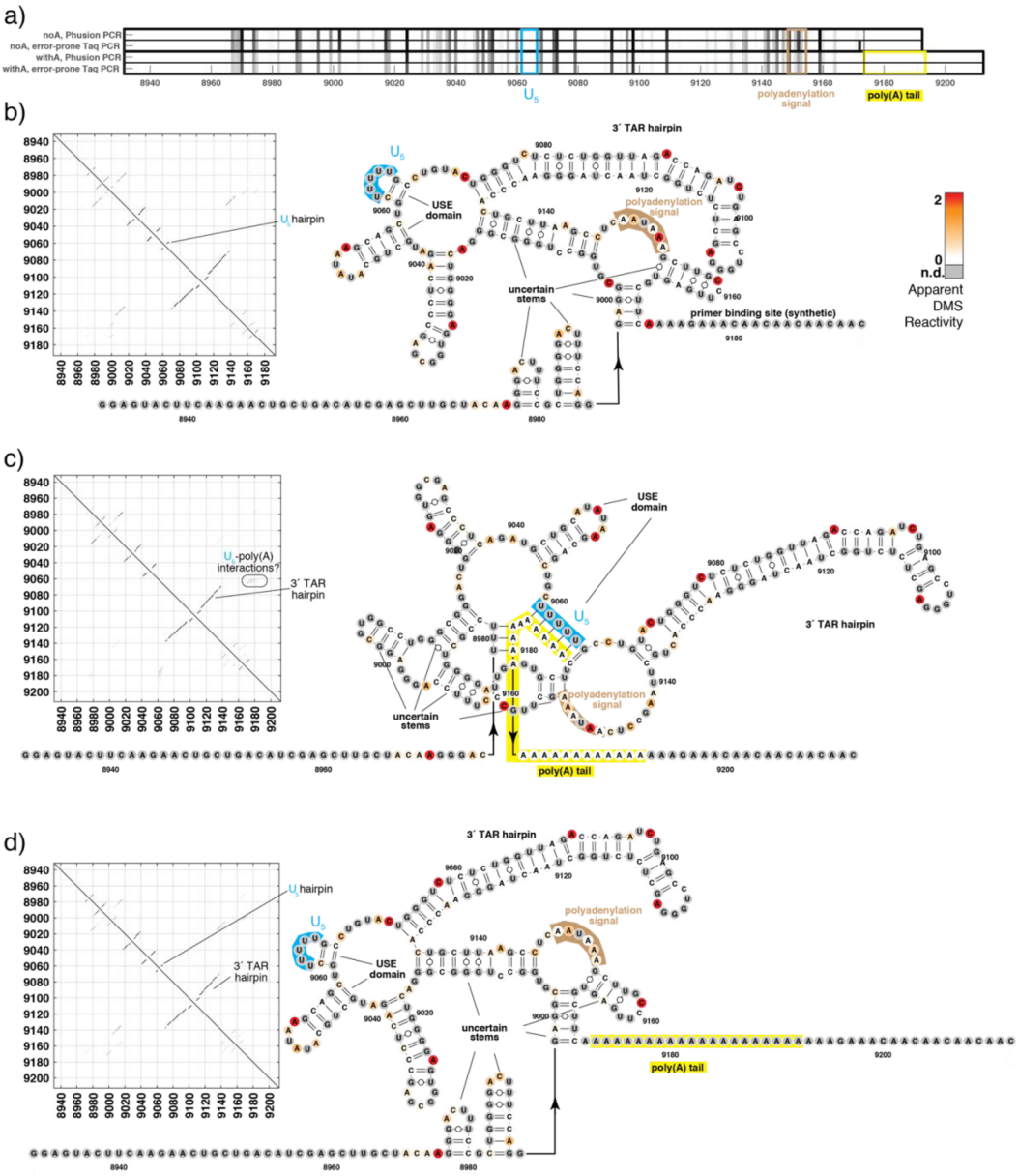


Figure 8. Reverse transcriptase read-through of DMS-modified nucleobases leads to anomalous reactivity data that can confound structural modeling of the HIV 3′ UTR by chemical mapping. (a) DMS profiles without and with the A_{20} tail are nearly indistinguishable. Reverse transcription through the poly(A) tail of the HIV 3′ UTR demonstrates a striking loss of apparent DMS reactivity (white). (b–d) Predicted secondary structures of HIV 3′ UTR (b), without the A_{20} tail, (c) with the A_{20} tail and using the conventional assumption that DMS protections in the tail reflect the formation of structure, and (d) with the A_{20} tail but ignoring DMS protections (gray) due to invisibility of chemical modifications to the RTase readout (deletion/mismatches introduced by SuperScript II with Mn^{2+}). In panels b–d, insets give bootstrap confidence values for each inferred base pair.

from the m^1A installed at −5, with a particularly strong signal for TGIRT-III (13% misincorporation). However, for all the RTases tested, we observed negligible mismatches across from the m^1A installed at −10, compared to the unmodified control (Figure 7c and Figure S6).

Overall, the spectrum of deletion/insertion/mismatch events produced by the RTase depends on the location of the m^1A within the poly(A) tract. Interestingly, the m^1A installed at −10 gave rise to no deletion, insertion, or mismatch signature for all the RTases. This modification, which would be encountered by the RTase after it has reverse-transcribed through a nine-adenosine stretch, appears invisible to current reverse transcriptases, producing neither RT stops (Figure 6) or other cDNA sequence signatures (Figure 7).

We expected the complexity of the deletion/insertion/mismatch spectra to be problematic for current data analysis packages developed to analyze chemical modifications based on Illumina sequencing of cDNAs. Indeed, analysis of the data sets above, which should show insertions, deletions, or mismatches at the single m^1A sites, produced anomalous results. For the m^1A-5 substrate analyzed with each of the RTases, ShapeMapper discovers the modification at −5 but also produces a spurious "hit" at position −10 (Figure 7d and Figure S7). Interestingly, the more recently developed packages ShapeMapper 2[30] and RNA Framework[31] assign

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

the mutational events from the same data set in different ways, filtering out fewer reads but giving apparent modifications at other positions throughout the A tract (Figure 7d). Other misassignments arise for m^1A-10 in all three packages (Figure 7d) as well as experiments using distinct reverse transcriptases (Figure S7). These different profiles, each with incorrect features, arise from different filters and assumptions for which positions to assign deletions, insertions, and/or mismatches for cDNAs. We could not find any reverse transcriptases or analysis package settings that recovered the expected profile of a single strong modification at the chemically installed 1-methyl-adenosine within the polyadenosine stretch.

Implications for RNA Structure Inference. The observation that RTases can produce complex cDNA products when bypassing modifications in polyadenosine stretches raises concerns for chemical mapping experiments used to biochemically probe RNA structure. After carrying out the studies above, we re-examined the data we had been acquiring on the HIV 3′ untranslated region (3′ UTR). This viral RNA segment is polyadenylated during its biogenesis *in vivo* and is hypothesized to exhibit functionally important structures, which may interconvert during different viral stages and recruit distinct partners.[12,18,54] We had originally seen paradoxical results when carrying out DMS mutational profiling[11] and mutate-and-map-seq experiments[29] without and with a poly(A) tail. On the one hand, we observed almost no difference across the entire 3′ UTR's DMS-MaP-seq profile in constructs without and with the poly(A) tail (Figure 8a). Here, the poly(A) tail was represented by a 20 adenosine stretch placed after the 3′ UTR polyadenylation site (the total stretch length is 24 adenosines due to an additional A at the polyadenylation site and three A's from our appended primer binding site). The DMS profile comparison (obtained by reverse transcription with SuperScript II with Mn^{2+}) suggested that there is a negligible structural effect from polyadenylation, at least for RNA probed *in vitro*.

Our original structure modeling efforts suggested a different picture. Modeling based on RNAstructure's *Fold* executable[55] guided by the DMS and mutate-and-map-seq data[29] (Figure S8) resulted in the secondary structures shown in Figure 8b,c. In the structure model without the poly(A) tail, stems corresponding to three USE (upstream sequence domain) hairpins and the long TAR (trans-activation response) element were modeled with high confidence (based on bootstrap uncertainty estimation;[29,35,56] see inset of Figure 8b and Figure S8). However, the inclusion of the poly(A) sequence resulted in a rearrangement in which one of the USE hairpins (sequence U_5, blue in Figure 8) shifted to the base pair with five of the introduced A's (yellow, Figure 8b), in contradiction with the near indistinguishability of the DMS profile at and around that region (Figure 8a).

The contradiction is resolved by the findings described above that chemical modifications within poly(A) are essentially invisible to reverse transcription. In accord with this picture, we observed negligible apparent DMS reactivity in the poly(A) tail appended to the HIV 3′ UTR. Any mismatches in the last three A's were masked by the use of a reverse transcription primer, including those A's, and we expected the RTase would bypass any chemical modifications at the 21 previous A's and, therefore, give no DMS signal there (Figure 6). Indeed, based on the analysis of the previous sections, we expected the RTase bypass would involve deletions in the cDNAs, and we confirmed this expectation with custom analysis scripts (Figure S9a); the resulting cDNAs are filtered out not only by our original analysis package ShapeMapper but also by ShapeMapper 2 and RNA Framework (Figure S9b). Given this bypass and the inability of current data processing frameworks to assign modifications within polyadenosine stretches (Figure S7), we reasoned that the best available option for structure modeling involves the use of *Fold* without any energetic bonuses or penalties in the poly(A) stretch, analogous to the treatment of G and U residues that do not give DMS signals in our conditions (gray nucleotides in Figure 8). Ignoring poly(A) data gives the model structure shown in Figure 8d. The poly(A) is modeled as fully unpaired, and this secondary structure recovers all of the base pairs of the poly(A)-less HIV 3′ UTR without any rearrangements. These results illustrate the importance of taking into account anomalous reverse transcription at poly(A) when interpreting chemical mapping data to infer RNA structures.

DISCUSSION

DMS and SHAPE mapping experiments give anomalous signatures at poly(A) stretches: a striking loss of reverse transcription cDNA termination products for adenosines that are at least six nucleotides upstream of a poly(A) 3′ end. Our results and data from previous studies show that a single nucleotide substitution in the poly(A) stretch abrogates this anomalous behavior. We have dissected each step of our structure mapping protocol to understand the mechanism of anomalous poly(A) DMS and SHAPE signatures. Experiments that attempted to "melt" a possible structure of poly(A) RNA or to displace it through base pairing failed to confirm any anomalous structure; NMR experiments also disfavored such an anomalous structure. Instead, the DMS and SHAPE signatures varied with the reverse transcriptases used to read the signatures, with some RTase enzymes showing nearly quantitative bypass of termination after reverse transcribing six adenosines and others shifting to bypass chemical modifications with even fewer adenosines. Analogous measurements on RNA substrates with m^1A installed at specific sites during chemical synthesis confirm that RTases bypass rather than terminate or introduce cDNA mutations at the modifications. The efficiency of bypass depends sharply on the number of A's that the RTase has already polymerized and on which RTase is used. Some recently engineered RTases (SuperScript IV, TGIRT-III, MarathonRT) bypass chemical modifications after polymerizing as few as 2−3 adenosines.

The behavior we observe with long polyadenosine stretches is compatible with available knowledge of the enzymatic activities and structures of reverse transcriptases. For example, the series of SuperScript enzymes developed from MMLV have a terminal nucleotidyl transferase (TdT) activity that favors the addition of dT to cDNA/RNA hybrids,[57] although this activity has been suppressed through engineered mutations in the SuperScript enzymes. Classic studies on the HIV-1 RTase at poly(A) stretches revealed template-primer slippage of RTase, resulting in insertions or deletions[45,58] or template switching.[59] It is not yet clear if the near-quantitative bypass mode we observe, which is triggered by a specific length of poly(A) (six nucleotides for SuperScript III), involves the same molecular steps as the TdT activity or the slippage mechanisms previously described. In particular, the previous experiments did not test whether TdT or the observed slippage events have a sharp and extended dependence on poly(A) length. Such

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

studies will be necessary to understand the connections with the various enzymatic activities that have been previously uncovered and the bypass described here. Other papers have noted a dependence of reverse transcription on nucleotides immediately 3′ to the site of polymerization across a variety of RTases[7,21,60−63] but again have not systematically varied the lengths of the poly(A) stretches. Further experiments will be necessary to understand these potentially complex and sharp dependencies on 3′ sequence and length and on RTase type.

From a structural point of view, a dependence of RTase activity on a previously reverse-transcribed stretch of adenosines is plausible. As illustrated in Figure 9a, an RTase paused at a chemical modification would be sensitive to the structure of a previously polymerized rA/dT hybrid duplex. Such a readout of the hybrid structure might enable the bypass through the roadblock, e.g., by the repositioning of the RNA template to bring the next (unmodified) ribonucleotide into the active site. Consistent with this model, crystallographic studies of the HIV-1 reverse transcriptase show that the RNA/cDNA hybrid is held within the RTase for numerous base pairs beyond the site of cDNA polymerization (Figure 9b), and the RTase active site domains are well-positioned to engage with the sequence of the RNA/cDNA hybrid that is polymerized immediately prior to the RNA in the active site. Other RTase domains such as the RNase H domain are more distal from the polymerization active site but can indirectly readout changes in the shape of the hybrid near the active site. Indeed, biochemical studies have shown that contacts of the RNA/cDNA hybrid with the HIV-1 RTase enzyme's RNase H domain (∼18 bp away from the polymerase active site) can be modulated by changes in the hybrid sequence.[65,66] Furthermore, crystallographic studies have revealed that the docking mode of the RNA/cDNA hybrid within the RTase changes with the hybrid sequence (compare subpanels of Figure 9b).[2462] In future mechanistic work, bringing a long (greater than 6 bp) rA/dT hybrid into the polymerase active site may help stabilize the "bypass" mode of reverse transcriptases and enable incisive kinetic and structural studies (Figure 9b, bottom-right subpanel).

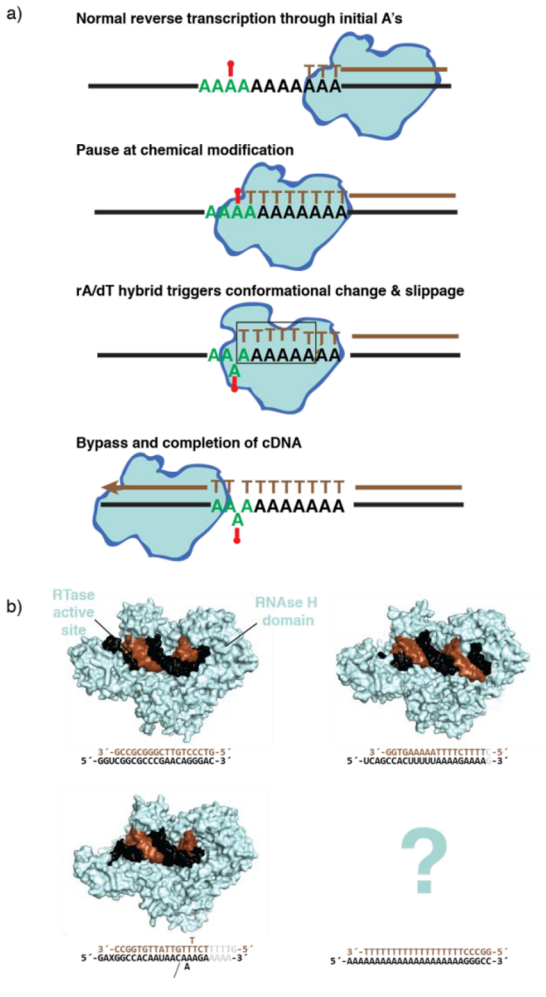


Figure 9. Possible mechanism for anomalous reverse transcription bypass at chemically modified poly(A) substrates. (a) Reverse transcriptase enzyme (cyan) polymerizing DNA (brown) on an RNA template (black) bypasses chemical modifications at which it would normally terminate (red) if it has previously reverse-transcribed adenosines. The mechanism may involve recognition of the poly rA-dT hybrid duplex (green) by the enzyme. (b) RNA/cDNA hybrids of different sequences show different docking modes to the HIV-1 reverse transcriptase (from bottom-left, clockwise, PDB IDs: 6BSI,[24] 4PQU,[64] 1HYS[65]). To aid visual comparison of differences, only the first 18 RNA and DNA nucleotides 3′ and 5′ to the active site, respectively, and any 5′ RNA overhangs are shown in the 3D models. Other base pairs are denoted in gray at the sequence depiction at the bottom of each panel. RNA = black. DNA = brown. Reverse transcriptase = cyan.

The biological significance of bypass of poly(A) chemical modifications is not yet clear, but retroviral RNA genomes and their functionally important cleavage products are polyadenylated; the genomes also harbor functionally important polypurine stretches (although these tracts have both G and A).[65−67] We speculate that the RTase bypass of chemical modifications at these sites may be important for the retrovirus to evade host responses involving chemical modification of viral RNAs. Our study has highlighted the same anomalous bypass effect in group II-intron-derived, rather than retroviral, reverse transcriptases. It is unclear whether these enzymes, which are also parts of selfish elements,[68] also encounter chemically modified poly(A) sequences during their evolution.

In addition to biological implications, our study has implications for two classes of biotechnology methods that involve the detection of RNA chemical modifications. First, given our results on the HIV 3′ UTR DMS mapping and structural inference, 3′ sequence context effects may be introducing systematic biases in structural studies based on chemical mapping using, e.g., the SHAPE-MaP and DMS-MaP-seq approaches. Second, 3′ sequence context effects may be causing m^1A and chemical modifications in natural mRNA samples to be bypassed by RTases and therefore undercounted, perhaps severely so in poly(A) tails. For both applications of reverse transcriptases, correcting these biases will require detailed characterization of sequence contexts of reverse transcription termination vs read-through, extending up to six nucleotides 3′ of the modified nucleotides.

CONCLUSION

We investigated the origins of a recently reported loss of chemical modification signal within unstructured polyadenosine stretches treated with SHAPE and DMS reagents. Experiments based on high-throughput sequencing, capillary electrophoresis, NMR spectroscopy, and specially designed RNA substrates implicate a previously unreported activity of reverse transcriptases: the enzymes bypass chemical mod-

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

ifications within polyadenosine stretches with near-quantitative efficiency. All tested natural and engineered reverse transcriptases exhibit this anomalous bypass behavior. Our study suggests polyadenosine-containing substrates that may allow incisive kinetic and structural dissection of how reverse transcriptase bypasses chemical modification. In addition, the results highlight potential biases in current methods for RNA structure mapping and detection of natural chemical modifications that merit further exploration.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.biochem.0c00020.

Reverse transcriptase conditions tested for poly(A) bypass; reverse transcription primers and ligation adapters; apparent DMS signals for nearly all 5-mer sequences read out through mutational profiling; capillary electropherograms; NMR spectra; Illumina sequencing; templates with additional designed base pairs further implicate the RTase bypass model; profiling deletions, insertions, and substitutions introduced by different reverse transcriptases across from m1 A in poly(A) stretches; differences in output of ShapeMapper, ShapeMapper 2, and RNA Framework analysis packages at 1-methyladenosine in the chemically synthesized Syn41 substrate; mutate-and-map-seq experiments on HIV-1 3′-UTR; supplemental analysis (PDF)

Primers and sequences (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Rhiju Das − *Department of Biochemistry, Stanford University School of Medicine, Stanford, California 94305, United States; Biophysics Program and Department of Physics, Stanford University, Stanford, California 94305, United States;* orcid.org/0000-0001-7497-0972; Email: rhiju@stanford.edu

Authors

Wipapat Kladwang − *Department of Biochemistry, Stanford University School of Medicine, Stanford, California 94305, United States*

Ved V. Topkar − *Biophysics Program, Stanford University, Stanford, California 94305, United States*

Bei Liu − *Department of Biochemistry, Duke University School of Medicine, Durham, North Carolina 27710, United States*

Ramya Rangan − *Biophysics Program, Stanford University, Stanford, California 94305, United States*

Tracy L. Hodges − *Biophysics Program, University of Michigan, Ann Arbor, Michigan 48109, United States*

Sarah C. Keane − *Biophysics Program and Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109, United States*

Hashim al-Hashimi − *Department of Biochemistry and Department of Chemistry, Duke University School of Medicine, Durham, North Carolina 27710, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.biochem.0c00020

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Eterna participants J. Anderson-Lee, E. Fisker, R. Wellington-Oguri, M. Wiley, and M. Zada, and S. Rouskin (MIT) for discussions and K. Liu and J. Yesselman for help with pilot studies and analysis. We thank A. Pyle (Yale University) for the gift of the MarathonRT enzyme. This study was funded by the National Institutes of Health through grants R01 GM100953, R35 GM122579, R21 CA219847, R21 AI145647, and P50 GM103297.

■ REFERENCES

(1) Atkins, J. F., Gesteland, R. F., and Cech, T. (2011) *RNA worlds: from life's origins to diversity in gene regulation*, p 361, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

(2) Li, X., Xiong, X., and Yi, C. (2017) Epitranscriptome sequencing technologies: decoding RNA modifications. *Nat. Methods 14* (1), 23−31.

(3) Dominissini, D., Nachtergaele, S., Moshitch-Moshkovitz, S., Peer, E., Kol, N., Ben-Haim, M. S., Dai, Q., Di Segni, A., Salmon-Divon, M., Clark, W. C., Zheng, G., Pan, T., Solomon, O., Eyal, E., Hershkovitz, V., Han, D., Dore, L. C., Amariglio, N., Rechavi, G., and He, C. (2016) The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA. *Nature 530* (7591), 441−446.

(4) Grozhik, A. V., and Jaffrey, S. R. (2017) Epitranscriptomics: Shrinking maps of RNA modifications. *Nature 551* (7679), 174−176.

(5) Darnell, R. B., Ke, S., and Darnell, J. E., Jr. (2018) Pre-mRNA processing includes N(6) methylation of adenosine residues that are retained in mRNA exons and the fallacy of "RNA epigenetics". *RNA 24* (3), 262−267.

(6) Zhao, B. S., Nachtergaele, S., Roundtree, I. A., and He, C. (2018) Our views of dynamic N(6)-methyladenosine RNA methylation. *RNA 24* (3), 268−272.

(7) Zhou, H., Rauch, S., Dai, Q., Cui, X., Zhang, Z., Nachtergaele, S., Sepich, C., He, C., and Dickinson, B. C. (2019) Evolution of a reverse transcriptase to map N(1)-methyladenosine in human messenger RNA. *Nat. Methods 16*, 1281−1288.

(8) Safra, M., Sas-Chen, A., Nir, R., Winkler, R., Nachshon, A., Bar-Yaacov, D., Erlacher, M., Rossmanith, W., Stern-Ginossar, N., and Schwartz, S. (2017) The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature 551* (7679), 251−255.

(9) Grozhik, A. V., Olarerin-George, A. O., Sindelar, M., Li, X., Gross, S. S., and Jaffrey, S. R. (2019) Antibody cross-reactivity accounts for widespread appearance of m(1)A in 5′UTRs. *Nat. Commun. 10* (1), 5126.

(10) Strobel, E. J., Yu, A. M., and Lucks, J. B. (2018) High-throughput determination of RNA structures. *Nat. Rev. Genet. 19* (10), 615−634.

(11) Zubradt, M., Gupta, P., Persad, S., Lambowitz, A. M., Weissman, J. S., and Rouskin, S. (2017) DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Methods 14* (1), 75−82.

(12) Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A., and Weeks, K. M. (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods 11* (9), 959−965.

(13) Yesselman, J. D., Tian, S., Liu, X., Shi, L., Li, J. B., and Das, R. (2018) Updates to the RNA mapping database (RMDB), version 2. *Nucleic Acids Res. 46* (D1), D375−D379.

(14) Wellington-Oguri, R., Fisker, E., Zada, M., Wiley, M., Townley, J., and Players, E. (2018) Evidence of an Unusual Poly(A) RNA Signature Detected by High-throughput Chemical Mapping. *bioRxiv*, 281147. (https://doi.org/10.1101/281147); Wellington-Oguri, R., Fisker, E., Zada, M., Wiley, M., Townley, J., and Players, E. (2020)

Evidence of an Unusual Poly(A) RNA Signature Detected by High-throughput Chemical Mapping. *Biochemistry*.
(15) Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M. (2005) RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc. 127* (12), 4223−31.
(16) Sarkar, N. (1997) Polyadenylation of mRNA in prokaryotes. *Annu. Rev. Biochem. 66*, 173−197.
(17) Colgan, D. F., and Manley, J. L. (1997) Mechanism and regulation of mRNA polyadenylation. *Genes Dev. 11* (21), 2755−2766.
(18) Das, A. T., Klaver, B., and Berkhout, B. (1999) A hairpin structure in the R region of the human immunodeficiency virus type 1 RNA genome is instrumental in polyadenylation site selection. *J. Virol. 73* (1), 81−91.
(19) Orelle, C., Carlson, E. D., Szal, T., Florin, T., Jewett, M. C., and Mankin, A. S. (2015) Protein synthesis by ribosomes with tethered subunits. *Nature 524* (7563), 119−124.
(20) Palka, C., Forino, N., Hentschel, J., Das, R., and Stone, M. D. (2019) Folding heterogeneity in the essential human telomerase RNA three-way junction. *bioRxiv*, 2019.12.14.876565v1.
(21) Hauenschild, R., Tserovski, L., Schmid, K., Thuring, K., Winz, M. L., Sharma, S., Entian, K. D., Wacheul, L., Lafontaine, D. L., Anderson, J., Alfonzo, J., Hildebrandt, A., Jaschke, A., Motorin, Y., and Helm, M. (2015) The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Res. 43* (20), 9950−64.
(22) Koutmou, K. S., Schuller, A. P., Brunelle, J. L., Radhakrishnan, A., Djuranovic, S., and Green, R. (2015) Ribosomes slide on lysine-encoding homopolymeric A stretches. *eLife 4*, e05534.
(23) Koscielniak, D., Wons, E., Wilkowska, K., and Sektas, M. (2018) Non-programmed transcriptional frameshifting is common and highly RNA polymerase type-dependent. *Microb. Cell Fact. 17* (1), 184.
(24) Tian, L., Kim, M. S., Li, H., Wang, J., and Yang, W. (2018) Structure of HIV-1 reverse transcriptase cleaving RNA in an RNA/DNA hybrid. *Proc. Natl. Acad. Sci. U. S. A. 115* (3), 507−512.
(25) Kladwang, W., Mann, T. H., Becka, A., Tian, S., Kim, H., Yoon, S., and Das, R. (2014) Standardization of RNA chemical mapping experiments. *Biochemistry 53* (19), 3063−3065.
(26) Tian, S., Yesselman, J. D., Cordero, P., and Das, R. (2015) Primerize: automated primer assembly for transcribing non-coding RNA domains. *Nucleic Acids Res. 43* (W1), W522−W526.
(27) Yoon, S., Kim, J., Hum, J., Kim, H., Park, S., Kladwang, W., and Das, R. (2011) HiTRACE: high-throughput robust analysis for capillary electrophoresis. *Bioinformatics 27* (13), 1798−1805.
(28) Seetin, M. G., Kladwang, W., Bida, J. P., and Das, R. (2014) Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol. *Methods Mol. Biol. 1086*, 95−117.
(29) Cheng, C. Y., Kladwang, W., Yesselman, J. D., and Das, R. (2017) RNA structure inference through chemical mapping after accidental or intentional mutations. *Proc. Natl. Acad. Sci. U. S. A. 114* (37), 9876−9881.
(30) Busan, S., and Weeks, K. M. (2018) Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA 24* (2), 143−148.
(31) Incarnato, D., Morandi, E., Simon, L. M., and Oliviero, S. (2018) RNA Framework: an all-in-one toolkit for the analysis of RNA structures and post-transcriptional modifications. *Nucleic Acids Res. 46* (16), e97.
(32) Macon, J. B., and Wolfenden, R. (1968) 1-Methyladenosine. Dimroth rearrangement and reversible reduction. *Biochemistry 7* (10), 3453−8.
(33) Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR 6* (3), 277−93.
(34) Goddard, T. D., and Kneller, D. G. *SPARKY 3*, University of California, San Francisco.
(35) Cordero, P., Kladwang, W., VanLang, C. C., and Das, R. (2012) Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry 51* (36), 7037−7039.
(36) Kwok, C. K., Tang, Y., Assmann, S. M., and Bevilacqua, P. C. (2015) The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends Biochem. Sci. 40* (4), 221−232.
(37) Rice, G. M., Leonard, C. W., and Weeks, K. M. (2014) RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA 20* (6), 846−854.
(38) Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W., Jr, Swanstrom, R., Burch, C. L., and Weeks, K. M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature 460* (7256), 711−716.
(39) Lavender, C. A., Lorenz, R., Zhang, G., Tamayo, R., Hofacker, I. L., and Weeks, K. M. (2015) Model-Free RNA Sequence and Structure Alignment Informed by SHAPE Probing Reveals a Conserved Alternate Secondary Structure for 16S rRNA. *PLoS Comput. Biol. 11* (5), e1004126.
(40) Dadonaite, B., Gilbertson, B., Knight, M. L., Trifkovic, S., Rockman, S., Laederach, A., Brown, L. E., Fodor, E., and Bauer, D. L. V. (2019) The structure of the influenza A virus genome. *Nat. Microbiol 4* (11), 1781−1789.
(41) Kutchko, K. M., Madden, E. A., Morrison, C., Plante, K. S., Sanders, W., Vincent, H. A., Cruz Cisneros, M. C., Long, K. M., Moorman, N. J., Heise, M. T., and Laederach, A. (2018) Structural divergence creates new functional features in alphavirus genomes. *Nucleic Acids Res. 46* (7), 3657−3670.
(42) Sas-Chen, A., and Schwartz, S. (2019) Misincorporation signatures for detecting modifications in mRNA: Not as simple as it sounds. *Methods 156*, 53−59.
(43) Kwok, C. K., Ding, Y., Sherlock, M. E., Assmann, S. M., and Bevilacqua, P. C. (2013) A hybridization-based approach for quantitative and low-bias single-stranded DNA ligation. *Anal. Biochem. 435* (2), 181−186.
(44) Simon, L. M., Morandi, E., Luganini, A., Gribaudo, G., Martinez-Sobrido, L., Turner, D. H., Oliviero, S., and Incarnato, D. (2019) In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. *Nucleic Acids Res. 47* (13), 7003−7017.
(45) Bebenek, K., Abbotts, J., Roberts, J. D., Wilson, S. H., and Kunkel, T. A. (1989) Specificity and mechanism of error-prone replication by human immunodeficiency virus-1 reverse transcriptase. *J. Biol. Chem. 264* (28), 16948−56.
(46) Sexton, A. N., Wang, P. Y., Rutenberg-Schoenberg, M., and Simon, M. D. (2017) Interpreting Reverse Transcriptase Termination and Mutation Events for Greater Insight into the Chemical Probing of RNA. *Biochemistry 56* (35), 4713−4721.
(47) Xu, H., Yao, J., Wu, D. C., and Lambowitz, A. M. (2019) Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. *Sci. Rep. 9* (1), 7953.
(48) Busan, S., Weidmann, C. A., Sengupta, A., and Weeks, K. M. (2019) Guidelines for SHAPE Reagent Choice and Detection Strategy for RNA Structure Probing Studies. *Biochemistry 58* (23), 2655−2664.
(49) Lee, J., Kladwang, W., Lee, M., Cantu, D., Azizyan, M., Kim, H., Limpaecher, A., Gaikwad, S., Yoon, S., Treuille, A., and Das, R. (2014) RNA design rules from a massive open laboratory. *Proc. Natl. Acad. Sci. U. S. A. 111* (6), 2122−2127.
(50) Gleghorn, M. L., Zhao, J., Turner, D. H., and Maquat, L. E. (2016) Crystal structure of a poly(rA) staggered zipper at acidic pH: evidence that adenine N1 protonation mediates parallel double helix formation. *Nucleic Acids Res. 44* (17), 8417−8424.
(51) Eichhorn, C. D., Feng, J., Suddala, K. C., Walter, N. G., Brooks, C. L., 3rd, and Al-Hashimi, H. M. (2012) Unraveling the structural complexity in a single-stranded RNA tail: implications for efficient ligand binding in the prequeuosine riboswitch. *Nucleic Acids Res. 40* (3), 1345−1355.

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

(52) Eichhorn, C. D., and Al-Hashimi, H. M. (2014) Structural dynamics of a single-stranded RNA-helix junction using NMR. *RNA 20* (6), 782−791.
(53) Zhao, C., Liu, F., and Pyle, A. M. (2018) An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA 24* (2), 183−195.
(54) Zarudnaya, M. I., Potyahaylo, A. L., Kolomiets, I. M., and Hovorun, D. M. (2013) Structural model of the complete poly(A) region of HIV-1 pre-mRNA. *J. Biomol. Struct. Dyn. 31* (10), 1044−1056.
(55) Reuter, J. S., and Mathews, D. H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinf. 11*, 129.
(56) Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011) Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry 50* (37), 8049−56.
(57) Chen, D., and Patton, J. T. (2001) Reverse transcriptase adds nontemplated nucleotides to cDNAs during 5′-RACE and primer extension. *BioTechniques 30* (3), 574−582. 582.
(58) Bebenek, K., Abbotts, J., Wilson, S. H., and Kunkel, T. A. (1993) Error-prone polymerization by HIV-1 reverse transcriptase. Contribution of template-primer misalignment, miscoding, and termination probability to mutational hot spots. *J. Biol. Chem. 268* (14), 10324−34.
(59) Buiser, R. G., Bambara, R. A., and Fay, P. J. (1993) Pausing by retroviral DNA polymerases promotes strand transfer from internal regions of RNA donor templates to homopolymeric acceptor templates. *Biochim. Biophys. Acta, Gene Struct. Expression 1216* (1), 20−30.
(60) Latham, G. J., and Lloyd, R. S. (1994) Deoxynucleotide polymerization by HIV-1 reverse transcriptase is terminated by site-specific styrene oxide adducts after translesion synthesis. *J. Biol. Chem. 269* (46), 28527−30.
(61) Forgacs, E., Latham, G., Beard, W. A., Prasad, R., Bebenek, K., Kunkel, T. A., Wilson, S. H., and Lloyd, R. S. (1997) Probing structure/function relationships of HIV-1 reverse transcriptase with styrene oxide N2-guanine adducts. *J. Biol. Chem. 272* (13), 8525−8530.
(62) Patterson, J. T., Nickens, D. G., and Burke, D. H. (2006) HIV-1 reverse transcriptase pausing at bulky 2′ adducts is relieved by deletion of the RNase H domain. *RNA Biol. 3* (4), 163−169.
(63) Tserovski, L., Marchand, V., Hauenschild, R., Blanloeil-Oillo, F., Helm, M., and Motorin, Y. (2016) High-throughput sequencing for 1-methyladenosine (m(1)A) mapping in RNA. *Methods 107*, 110−121.
(64) Das, K., Martinez, S. E., Bandwar, R. P., and Arnold, E. (2014) Structures of HIV-1 RT-RNA/DNA ternary complexes with dATP and nevirapine reveal conformational flexibility of RNA/DNA: insights into requirements for RNase H cleavage. *Nucleic Acids Res. 42* (12), 8125−8137.
(65) Sarafianos, S. G., Das, K., Tantillo, C., Clark, A. D., Jr., Ding, J., Whitcomb, J. M., Boyer, P. L., Hughes, S. H., and Arnold, E. (2001) Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. *EMBO J. 20* (6), 1449−1461.
(66) Kvaratskhelia, M., Budihas, S. R., and Le Grice, S. F. (2002) Pre-existing distortions in nucleic acid structure aid polypurine tract selection by HIV-1 reverse transcriptase. *J. Biol. Chem. 277* (19), 16689−16696.
(67) Coutsinos, D., Invernizzi, C. F., Moisi, D., Oliveira, M., Martinez-Cajas, J. L., Brenner, B. G., and Wainberg, M. A. (2011) A template-dependent dislocation mechanism potentiates K65R reverse transcriptase mutation development in subtype C variants of HIV-1. *PLoS One 6* (5), e20208.
(68) Lambowitz, A. M., and Zimmerly, S. (2011) Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harbor Perspect. Biol. 3* (8), a003616.

https://dx.doi.org/10.1021/acs.biochem.0c00020
Biochemistry XXXX, XXX, XXX−XXX

Supplemental Information for: Anomalous reverse transcription through chemical modifications in polyadenosine stretches

Wipapat Kladwang[1], Ved T. Topkar[2], Bei Liu[3], Ramya Rangan[2], Tracy L. Hodges[4], Sarah C. Keane[4,5], Hashim al-Hashimi[3,6], Rhiju Das[1,2,7,*]

[1]Department of Biochemistry, Stanford University School of Medicine, Stanford CA 94305
[2]Biophysics Program, Stanford University, Stanford CA 94305
[3]Department of Biochemistry, Duke University School of Medicine, Durham NC 27710
[4]Biophysics Program, University of Michigan, Ann Arbor MI 48109
[5]Department of Chemistry, University of Michigan, Ann Arbor MI 48109
[6]Department of Chemistry, Duke University School of Medicine, Durham NC 27710
[7]Department of Physics, Stanford University, Stanford CA 94305

*Corresponding author: rhiju@stanford.edu.

Table S1. Reverse transcriptase conditions tested for poly(A) bypass

No.	Reverse-transcriptase	Buffer*	Reaction conditions	Enzyme source	Part number	Stock	Units/reaction
1	SuperScript-II	25 mM Tris-HCl- pH 8.3, 75 mM KCl, 3 mM $MgCl_2$, 5 mM DTT**	42 °C	Thermo Fisher Scientific	18064071	200 U/μL	20 Units
2	SuperScript-II + Mn^{2+}	25 mM Tris-HCl- pH 8.3, 75 mM KCl, 6 mM $MnCl_2$, 5 mM DTT	42 °C	Thermo Fisher Scientific	18064071	200 U/μL	20 Units
3	SuperScript-III	25 mM Tris-HCl- pH 8.3, 75 mM KCl, 3 mM $MgCl_2$, 5 mM DTT**	48 °C	Thermo Fisher Scientific	18080085	200 U/μL	20 Units
4	SuperScript-IV	50 mM Tris-HCl- pH 8.3, 4 mM $MgCl_2$, 50 mM KCl, 10 mM DTT**	50 °C	Thermo Fisher Scientific	18090050	200 U/μL	20 Units
5	TGIRT-III	50 mM Tris-HCl-pH 8, 75 mM KCl, 3 mM $MgCl_2$ 5 mM DTT	57 °C	InGex, LLC	TGIRT50	280 U/μL	140 Units
6	Marathon RT	50 mM Tris-HCl- pH 8.5, 100 mM KCl, 2 mM $MgCl_2$ 10 mM DTT	42 °C	Courtesy of Dr. A. Pyle, Yale University	n/a	5 μM	4.6 pmol
7	AMV	25 mM Tris-HCl- pH 8.3, 100 mM KCl, 5 mM $MgCl_2$, 2 mM DTT**	42 °C	AMS Biotechnology Ltd.	AMS.AMV007-1	20 U/μL	20 Units
8	MMLV	50 mM Tris-HCl- pH 8.3, 75 mM KCl, 3 mM $MgCl_2$, 5 mM DTT**	42 °C	Thermo Fisher Scientific	AM2043	100 U/μL	20 Units

*60 min reactions in 5 μL volumes

**Manufacture-supplied buffer was used.

Table S2. Reverse transcription primers (RTB) and ligation adapters for TOD-S1 and TOD-S7 MAP-seq cDNA library preparation. See Supplemental File sequences.xlsx for sequences.

RNA	SS-II	SS-II + Mn	SS-III	SS-IV	TGIRT	Marathon	AMV	MMLV	Ligation adapter
TOD-S1, no modification	RTB000	RTB001	RTB002	RTB003	RTB004	RTB005	RTB006	RTB007	P_TruSeqAdapt01_p
TOD-S1, 1M7	RTB008	RTB009	RTB010	RTB011	RTB012	RTB013	RTB014	RTB015	P_TruSeqAdapt01_p
TOD-S1, NMIA	RTB016	RTB017	RTB018	RTB019	RTB020	RTB021	RTB022	RTB023	P_TruSeqAdapt01_p
TOD-S1, DMS	RTB024	RTB025	RTB026	RTB027	RTB028	RTB029	RTB030	RTB031	P_TruSeqAdapt01_p
TOD-S7, no modification	RTB000	RTB001	RTB002	RTB003	RTB004	RTB005	RTB006	RTB007	P_TruSeqAdapt02_p
TOD-S7, 1M7	RTB008	RTB009	RTB010	RTB011	RTB012	RTB013	RTB014	RTB015	P_TruSeqAdapt02_p
TOD-S7, NMIA	RTB016	RTB017	RTB018	RTB019	RTB020	RTB021	RTB022	RTB023	P_TruSeqAdapt02_p
TOD-S7, DMS	RTB024	RTB025	RTB026	RTB027	RTB028	RTB029	RTB030	RTB031	P_TruSeqAdapt02_p

Table S3. Reverse transcription primers (RTB) and ligation adapters for Syn41 MAP-seq cDNA library preparation. See Supplemental File sequences.xlsx for sequences.

RNA	SS-II	SS-II + Mn	SS-III	SS-IV	TGIRT	AMV	MMLV	Ligation adapter
Syn41 (no modification)	RTB000	RTB001	RTB002	RTB003	RTB004	RTB005	RTB006	pAadaptBp
m^1A-A5	RTB007	RTB008	RTB009	RTB010	RTB011	RTB012	RTB013	P_TruSeqAdapt01_p
m^1A-A10	RTB014	RTB015	RTB016	RTB017	RTB018	RTB019	RTB020	P_TruSeqAdapt02_p

Table S4. RTB barcoded reverse transcription primers for mutate-and-map-seq/DMS-MaP-seq experiments on HIV 3´-UTR RNAs. See Supplemental File sequences.xlsx for sequences.

Number	RNA*	RTB primer	
		No mod	High DMS
1	P4P6.noHP-Phu	RTB000	RTB001
2	P4P6.noHP-Taq	RTB002	RTB003
3	NL4-3 with polyA tail -Phu	RTB004	RTB005
4	NL4-3 with polyA tail -Taq	RTB006	RTB007
5	WO-A-Phu	RTB008	RTB009
6	WO-A.noHP-Taq	RTB010	RTB011

* Phu – standard PCR preparation with Phusion polyerase, Taq – error-prone PCR with Taq polymerase and Mn^{2+}, to introduce mutations for mutate-and-map-seq analysis, NL4-3 – HIV 3´ UTR sequence from NL4-3 genome, WO-A – no A_{20}.

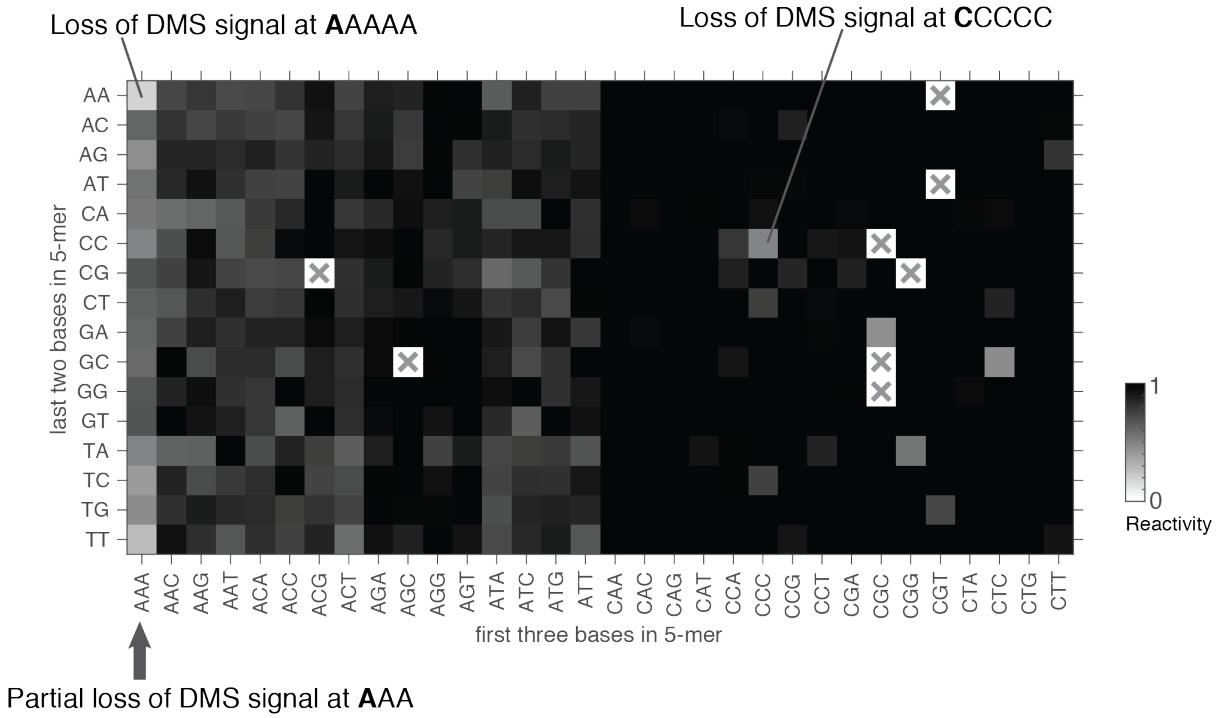


Figure S1. Apparent DMS signals for nearly all 5-mer sequences read out through mutational profiling.
Data are for denatured influenza A mRNAs averaged for each occurrence of the 5-mer. Reactivities are normalized to mean DMS reactivity at all 5-mers for which data are available. X's mark 5-mers for which data were not available. Loss of DMS signals at AAA triplets, and stronger loss at AAAAA and CCCCC are marked. Data reflect indels/mismatches introduced by TGIRT-III reverse transcription of DMS-modified RNA, read out by Illumina sequencing and analyzed by the RNA Framework package; see Simon, L. M.; Morandi, E.; Luganini, A.; Gribaudo, G.; Martinez-Sobrido, L.; Turner, D. H.; Oliviero, S.; Incarnato, D., In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. *Nucleic Acids Res* **2019,** *47* (13), 7003-7017.

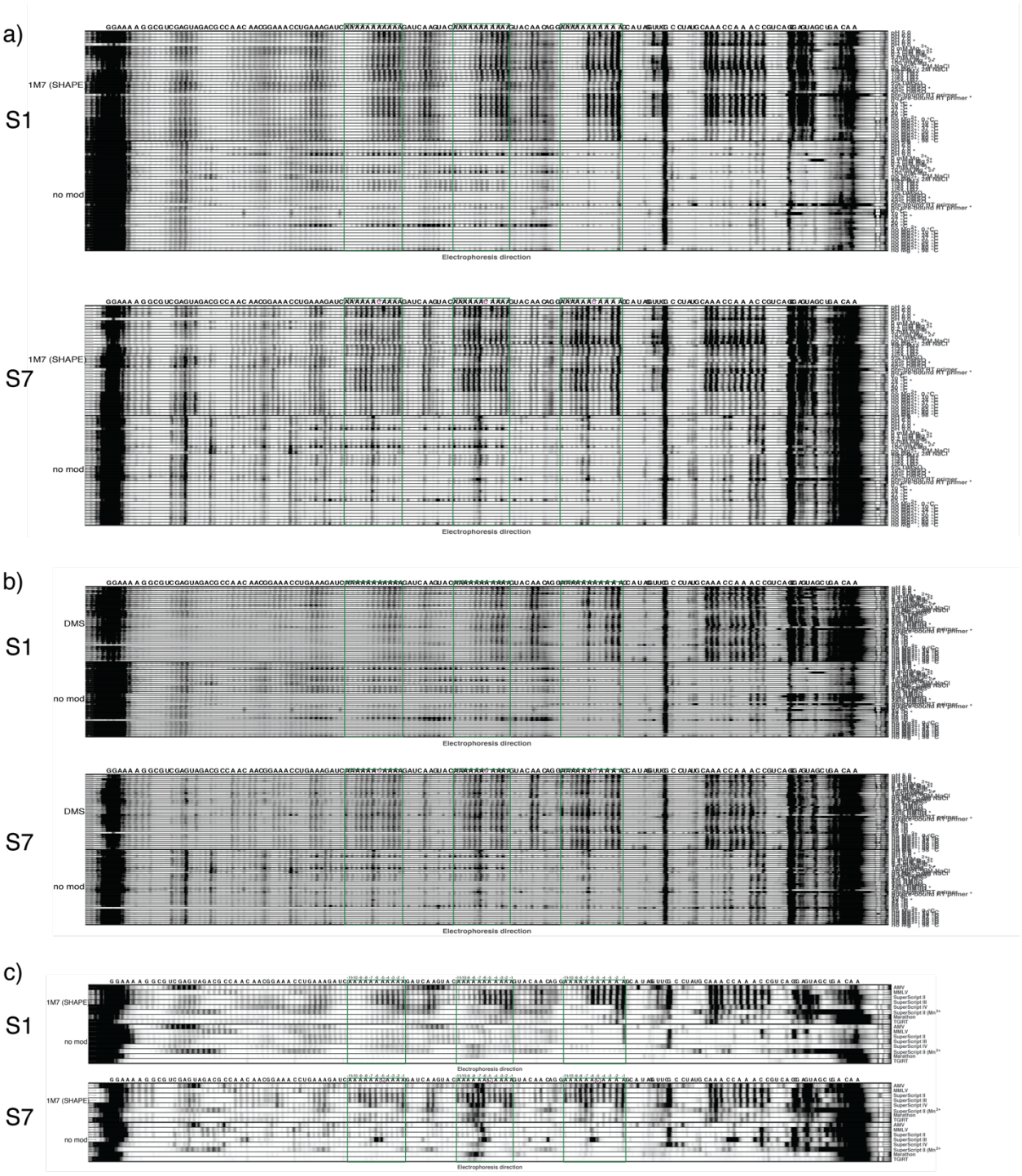


Figure S2. Capillary electropherograms for TOD-S1 and TOD-S7 experiments. (a,b) Variation of solution conditions during (a) SHAPE (1M7) or (b) DMS mapping, with no modification control data acquired side-by-side shown underneath. (c) Variation of reverse transcriptases after SHAPE (1M7) mapping. In each panel, stretches of A_{11} of AAAAAACAAAA are outlined in green boxes. Data reflect termination of reverse transcription by SuperScript III reverse transcriptase, read out by capillary electrophoresis.

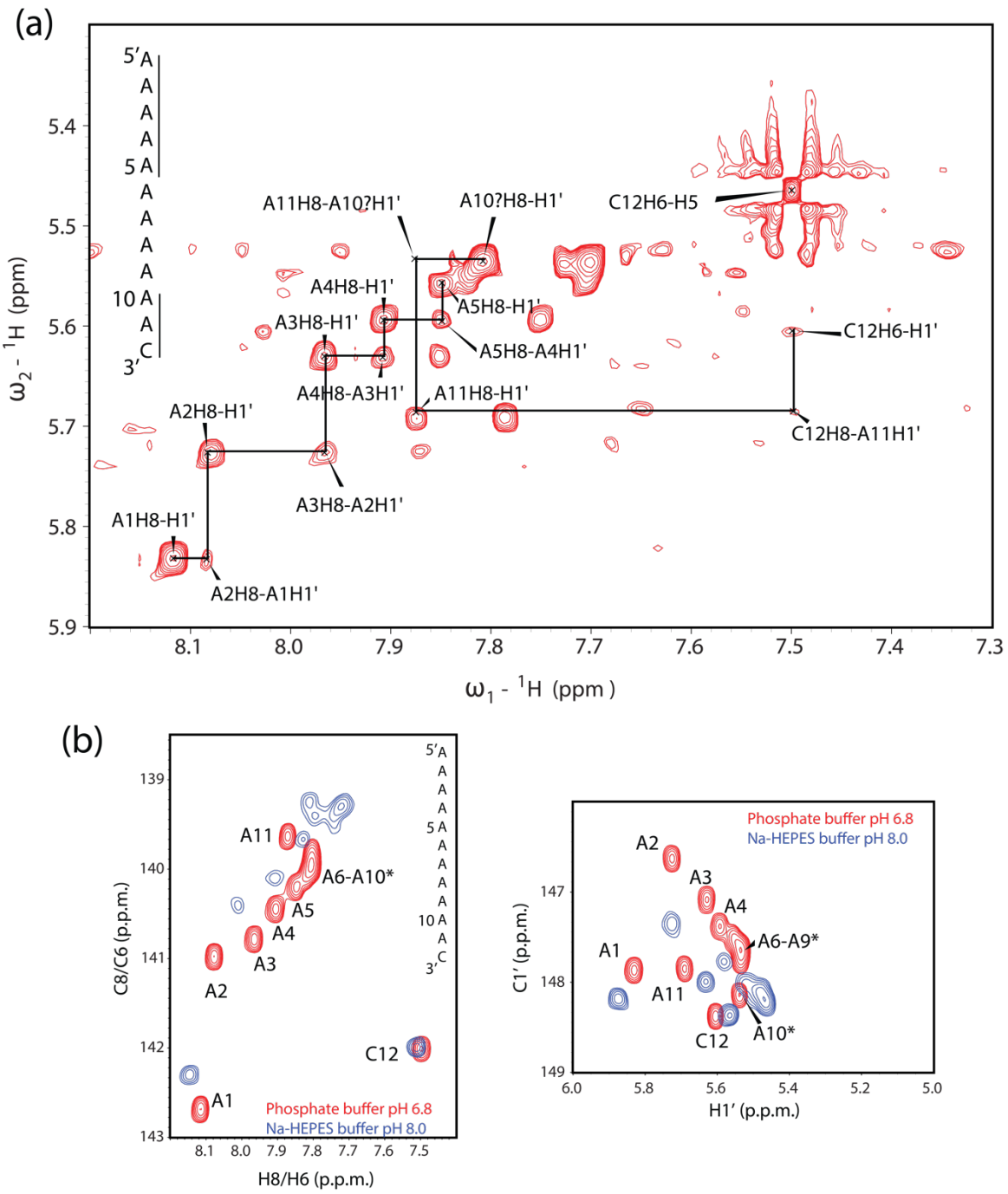


Figure S3. NMR spectra of $A_{11}C$ RNA for unusual structures. (a) 2D NOESY spectra of $A_{11}C$ RNA (5'-AAAAAAAAAAAC-3') in phosphate buffer (25 mM NaCl, 15 mM sodium phosphate, pH 6.8, 25°C; RNA concentration of 1 mM) enable chemical shift assignment. (b) 2D [^{13}C,^{1}H] HSQC spectra of C8 and H8 atoms within aromatic bases and C1′ and H1′ within riboses for A11C RNA in Na-HEPES buffer that matches our chemical mapping experiments (blue, 50 mM Na-HEPES, pH 8.0, 10 mM $MgCl_2$, T = 25°C; RNA concentration, 1 mM) and in phosphate buffer more standard for NMR (red). Chemical shifts and NOE cross-peaks remain consistent with a dominant A-form like structure for the RNA across buffer conditions.

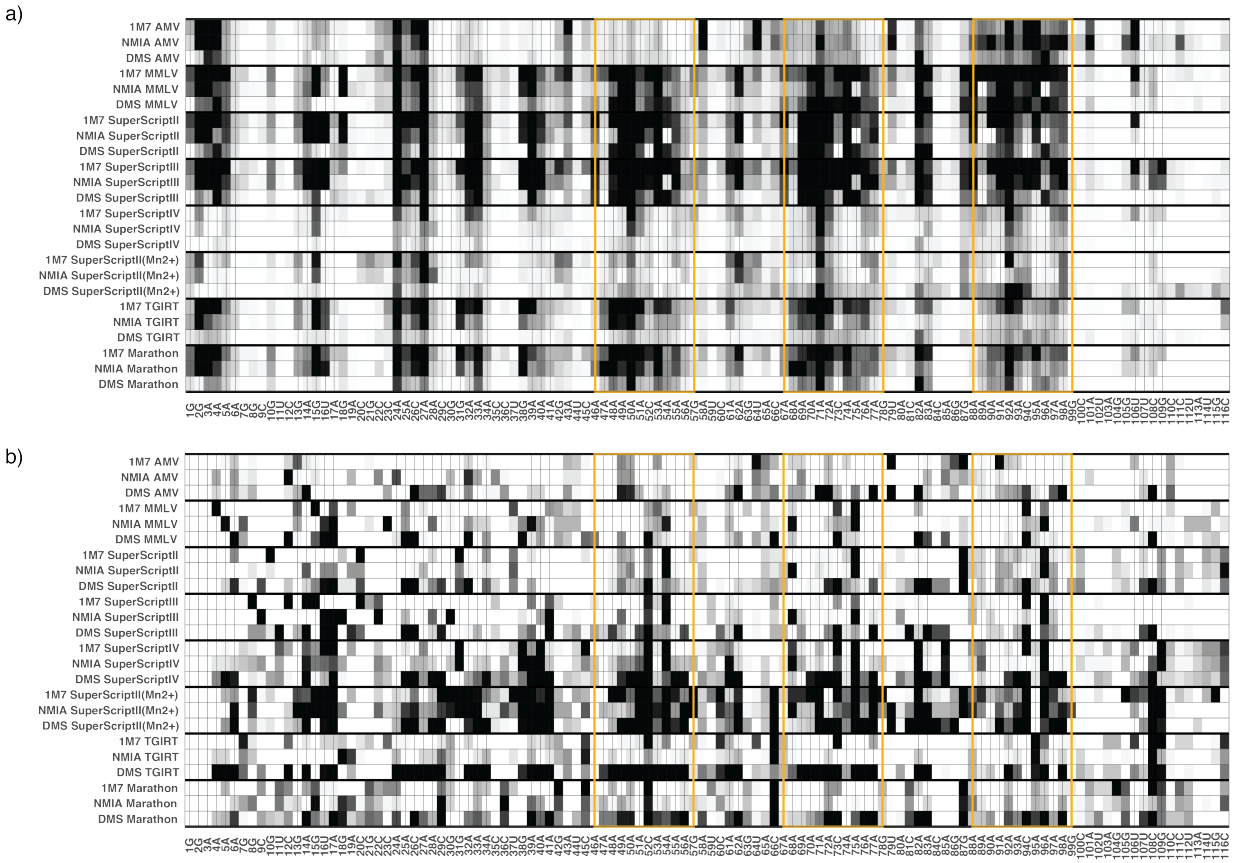


Figure S4. Illumina sequencing confirms bypass at poly(A) stretches depending on reverse transcriptase. Fractions of products (a) terminated or (b) inducing mismatches/deletions in TOD-S7 RNA, after background subtraction, for each of the reverse transcriptases specified in the axis labels. Thin black vertical lines mark bases designed to be unpaired in the molecules. Gold rectangles highlight repeats of AAAAAACAAAA stretch.

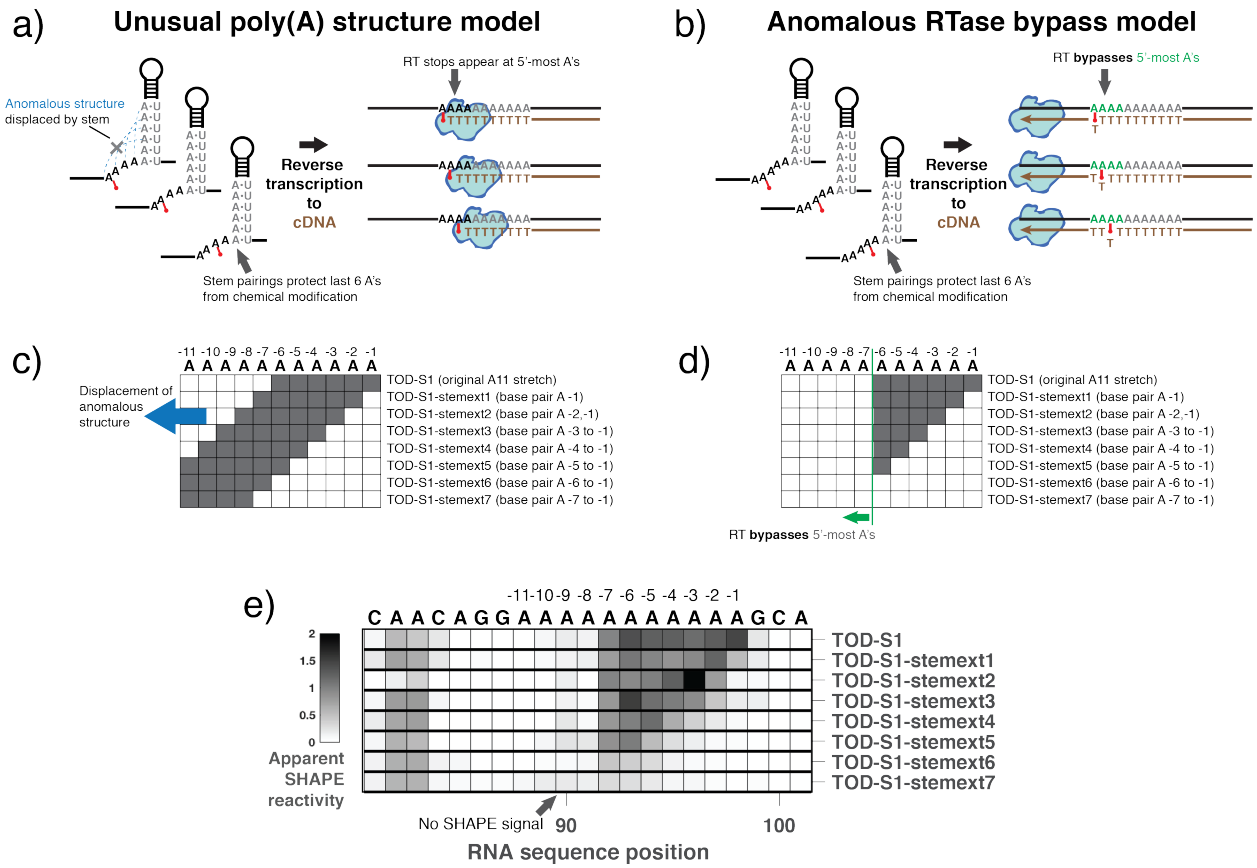


Figure S5. Templates with additional designed base pairs further implicate the RTase bypass model. (a-b) TOD-S1-ext RNAs introduce stretches of U's near each of the three 11-adenosine stretches and were designed to extend the RNA's three helical stems to include additional A-U pairs. (a,c) in the poly(A) structure model, if A's were taken out of an anomalous structure and into these canonical base pairs, the poly(A) anomalous structure (a) and associated chemical modification signals (c) should shift 5'. (b,d) In the 'reverse transcriptase bypass' model, the chemical modification signals at newly paired A's should disappear, but no new cDNA products corresponding to new chemical modifications at other A's should appear. (e) The experimental data match the predictions of the RTase bypass model (d) and disfavor the poly(A) structure model (c). Data reflect termination of reverse transcription by SuperScript III reverse transcriptase, read out by capillary electrophoresis.

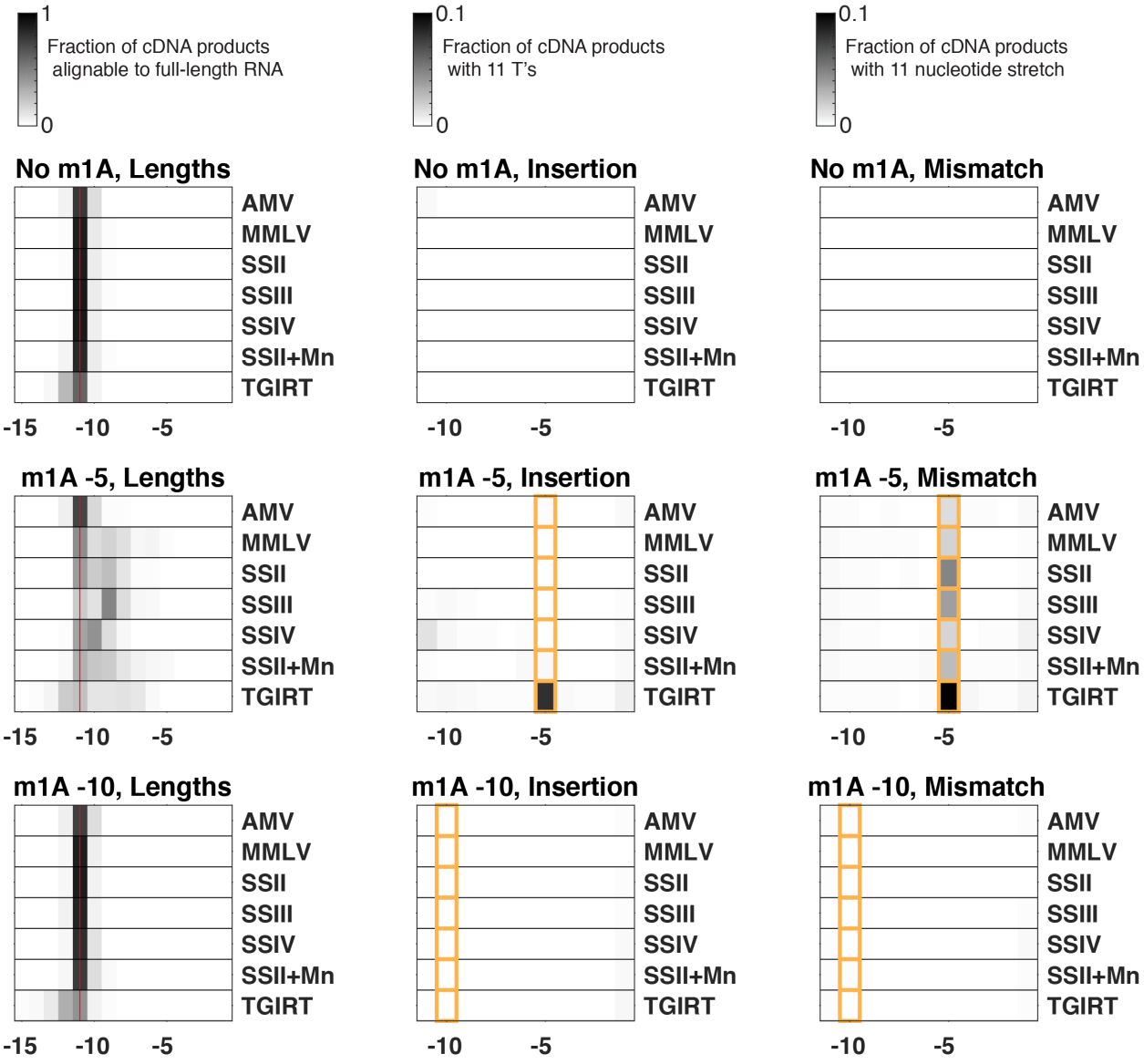


Figure S6. Profiling deletions, insertions, and substitutions introduced by different reverse transcriptases across from m^{1}A in poly(A) stretches. Data are for reverse transcription of chemical synthesized Syn41 RNA substrate without chemical modifications and with m^{1}A installed at -5 and -10 positions (highlighted with gold rectangles).

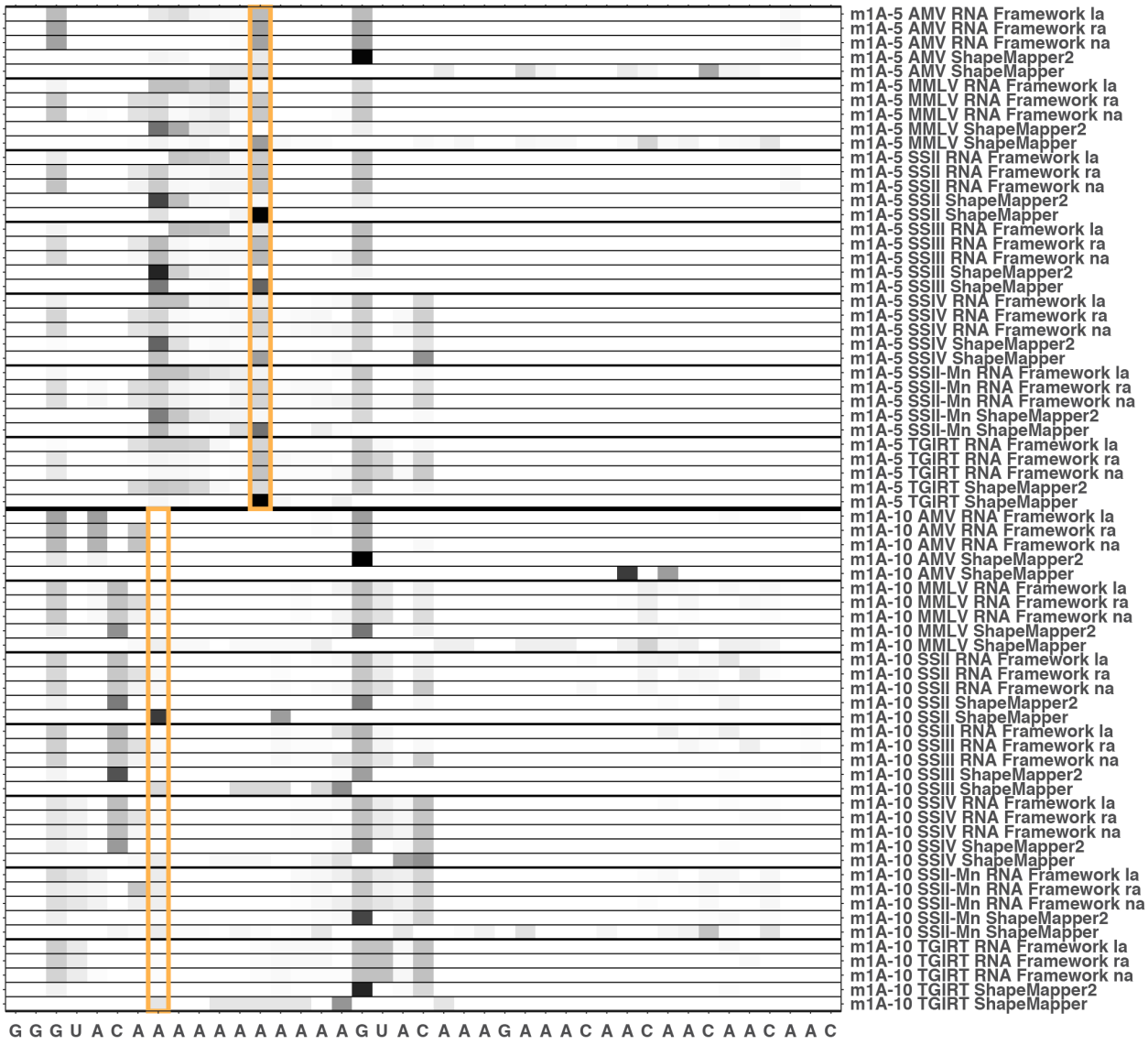


Figure S7. Differences in output of ShapeMapper, ShapeMapper 2, and RNA Framework analysis packages at 1-methyl-adenosine in the chemically synthesized Syn41 substrate. Analysis of reverse transcription cDNA products with single m^1A installed at -5 and -10 positions during chemical synthesis (highlighted with gold rectangles). Data from Syn41 substrates without m^1A were used for background subtraction. Data are normalized so that 1.0 corresponds to detection of an insertion, deletion, or mismatch in every cDNA by the package. Abbreviations: la – left assignment of ambiguous deletions (default for RNA Framework), ra – right assignment of ambiguous deletions, na – ignore ambiguous deletions.

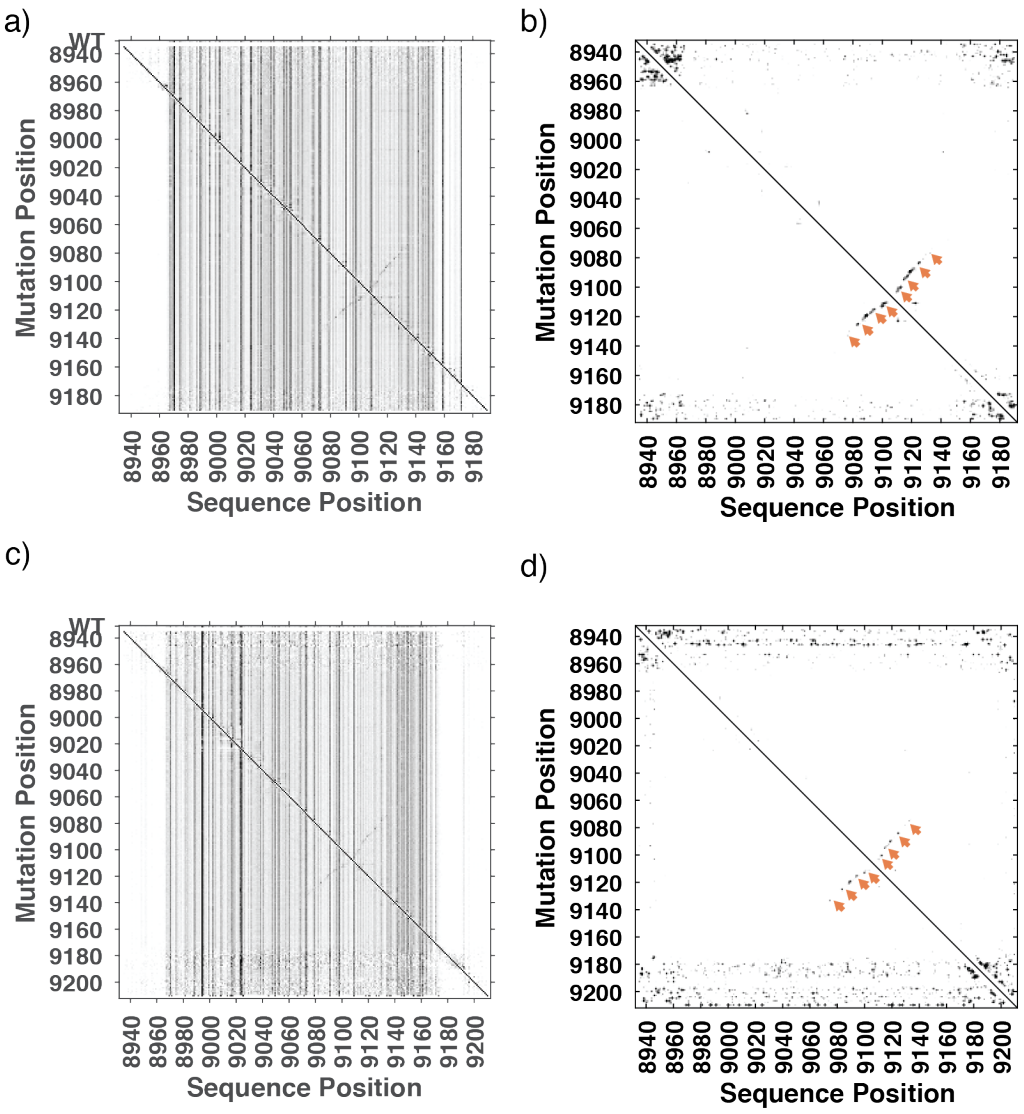


Figure S8. Mutate-and-map-seq experiments on HIV-1 3'-UTR. Numbering is for HIV-1 NL4-4 reference genome. Construct sequences end at polyadenylation site (a,b) or are further extended by $(A)_{20}$ (c,d); each are then extended by a 20-nt primer binding site. (a,c) Two-dimensional M2-seq signal shows mutation rate (primarily reflecting DMS reactivity) at each sequence position given detection of one mutation (primarily deletion/mismatches introduced by error-prone PCR) at other sites. (b,d) Z-score maps (smoothed) based on M2-seq signals; arrows mark stems that are automatically detected with the M2-net algorithm and are also clear visually. All four stems comprising the HIV-1 3' TAR hairpin are detected, in both data sets. Data reflect deletions and mismatches in cDNA assigned by Shape Mapper and M2seq.py pipeline. See Cheng, C.Y., Kladwang, W., Yesselman, J.D., and Das, R. (2017), "RNA structure inference through chemical mapping after accidental or intentional mutations", *Proceedings of the National Academy of Sciences U.S.A.* 114 (37) : 9876 – 9881.

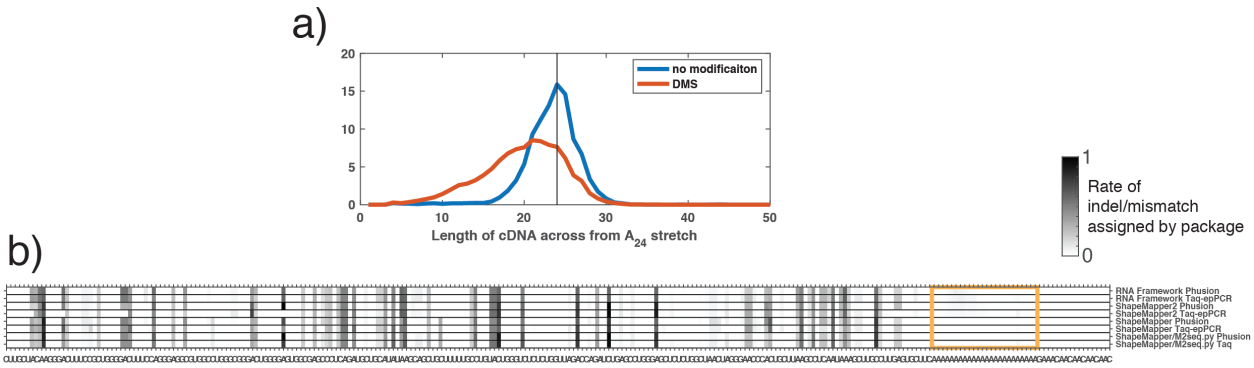


Figure S9. Supplemental analysis of DMS-MaP-seq of HIV 3´ UTR with polyadenosine tail. (a) Shift in lengths of cDNA poly(dT) regions reverse transcribed across from polyadenosine stretches of HIV 3´ UTR, in DMS-modified RNA compared to unmodified RNA. Reverse transcription was carried out with SuperScript II with Mn^{2+} to promote mutational bypass. (b) Mutational profiles inferred by ShapeMapper, ShapeMapper 2, and RNA Framework analysis packages assign mismatch/deletion events to different positions in polyadenosine tail (highlighted in gold rectangle), but all packages give apparently low reactivity. Data from unmodified RNA were used for background subtraction. Data are normalized as in Cheng, C.Y., Kladwang, W., Yesselman, J.D., and Das, R. (2017), "RNA structure inference through chemical mapping after accidental or intentional mutations", *Proceedings of the National Academy of Sciences U.S.A.* 114 (37): 9876 – 9881.