



HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2020 March 23.

Published in final edited form as:

Nat Methods. 2019 December ; 16(12): 1281–1288. doi:10.1038/s41592-019-0550-4.

Evolution of a Reverse Transcriptase to Map *N*¹-Methyladenosine in Human mRNA

Huiqing Zhou^{1,2}, Simone Rauch^{1,3}, Qing Dai^{1,4}, Xiaolong Cui^{1,2}, Zijie Zhang^{1,2,3,4}, Sigrid Nachtergaele^{1,2,4}, Caraline Sepich^{1,2,3,5}, Chuan He^{1,2,3,4,*}, Bryan C. Dickinson^{1,*}

¹Department of Chemistry, the University of Chicago, Chicago, Illinois 60637, USA

²Institute for Biophysical Dynamics, the University of Chicago, Chicago, Illinois 60637, USA

³Department of Biochemistry and Molecular Biology, the University of Chicago, Chicago, Illinois 60637, USA

⁴Howard Hughes Medical Institute, the University of Chicago, Chicago, Illinois 60637, USA

⁵University of Chicago Medical Scientist Training Program, Chicago, United States

Abstract

Chemical modifications on messenger RNA are increasingly recognized as a critical regulatory layer in the flow of genetic information, but quantitative tools to monitor RNA modifications in a whole-transcriptome and site-specific manner are lacking. Here we describe a versatile directed evolution platform that rapidly selects for reverse transcriptases that install mutations at sites of a given type of RNA modification during reverse transcription, allowing for site-specific identification of the modification. To develop and validate the platform, we evolved the HIV-1 reverse transcriptase against *N*¹-methyladenosine (m¹A). Iterative rounds of selection yielded reverse transcriptases with both robust read-through and high mutation rates at m¹A sites. The optimal evolved reverse transcriptase enabled detection of well-characterized m¹A sites and revealed hundreds of m¹A sites in human messenger RNA. Together, this work develops and validates the reverse transcriptase evolution platform, and provides new tools, analysis methods, and datasets to study m¹A biology.

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

*To whom correspondence should be addressed., chuanhe@uchicago.edu and dickinson@uchicago.edu.

AUTHOR CONTRIBUTIONS

B.C.D., C.H., and H.Z. conceived the ideas for the project and designed experiments. H.Z. performed the directed evolution experiments with assistance from S.R. performing the initial screening experiment on the Broccoli DNA. H.Z. prepared protein samples, and carried out biochemical characterizations, NGS library preparations, NGS data processing and analyses. Q.D. synthesized RNA oligonucleotides used in this study. S.N. and C.S. assisted with cell culturing for biological RNA library preparations. X.C. and Z.Z. assisted with NGS data analysis pipeline, enrichment and statistical analyses. H.Z., C.H., and B.C.D. wrote the manuscript with critical inputs from S.R., Q.D., X.C. Z.Z., S.N., and C.S.

COMPETING FINANCIAL INTERESTS

H.Z., C.H., and B.C.D. have filed a provisional patent application for the RTs described in this manuscript. C.H. is a scientific founder and a member of the scientific advisory board of Accent Therapeutics, Inc.

Editor's summary

Robust m¹A mapping of the human transcriptome is enabled by directed evolution of an HIV-1 RT with robust read-through and high mutation rates at m¹A sites along with development of new tools for data analysis.

INTRODUCTION

Messenger RNA (mRNA) modifications play critical regulatory roles in a range of mammalian biological processes, such as cell differentiation, sex determination and development^{1,2}, and are dysregulated in a number of human diseases³. RNA modifications such as *N*⁶-methyladenosine (m⁶A) and *N*¹-methyladenosine (m¹A) have been mapped on a transcriptome-wide scale by antibody-based enrichment methods (e.g. m⁶A-seq⁴, MeRIP-Seq⁵, m¹A-seq⁶ and m¹A-ID-seq^{7,8}). While enrichment-based approaches are powerful for outlining broad distributions of modifications, such profiling methods generally require large sample quantities and do not provide single-base resolution or quantitative information of the modification stoichiometry, limiting their utility in many biological experiments.

Single-base resolution techniques that allow detection of the exact location and the modification stoichiometry on a transcript are rapidly being developed^{7,9–12}. Base-resolution methods using next-generation sequencing (NGS) typically rely on reading signatures (i.e. mutations or stops) specifically introduced at the modification site by a reverse transcriptase (RT). Notably, mutations are more advantageous over stops as the latter is subjected to background noise arising from unintended stops at secondary or tertiary RNA structures, fragmentation biases, and RNA degradation.

There has been growing interest in engineering RTs for various applications in NGS. For examples, RTs were engineered to facilitate read-through of RNA secondary structures with improved thermostability^{13,14}, repurposed from DNA polymerases to harness their proof-reading activity¹⁵, and evolved to recognize modifications such as *N*⁵-methyldeoxycytosine¹⁶ and m⁶A¹⁷. However, few engineered RTs were applied successfully in transcriptome-wide mapping of targeted modifications. We reasoned that this methodological paucity can be attributed to limitations in current cumbersome and low-throughput RT selection strategies. Additionally, there are no high-throughput assays that directly detect RT signatures at modified RNA, which hinders the identification of enzymes with the desired properties for applications in NGS^{16,17}.

To develop a versatile and high-throughput platform to evolve RTs for RNA modifications, we focused on m¹A, which carries a positive charge and strongly disrupts canonical Watson-Crick-Franklin base-pairing (Fig. 1a). Although m¹A has long been known to be present in transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs)^{10,18–20}, mass spectrometry and transcriptome-wide m¹A profiling studies^{6,8} revealed that m¹A also occurs in mRNA and long non-coding RNA (lncRNA). Recently, two base-resolution mapping studies of m¹A^{11,21} offered contradictory findings regarding the number of reported m¹A sites in human mRNA^{6,8,22}. Among many reasons for such discrepancies, the TGIRT¹³ used to read through m¹A exhibits low activity in primer incorporation¹⁴. Such low RT activity may render it insensitive and incapable of generating reproducible m¹A maps, especially for m¹A of relatively low abundance in mRNA. We viewed these challenges in mapping the mammalian m¹A methylome as an opportunity to develop new general approaches to create robust RTs for more accurate and quantitative readouts of RNA modifications.

Here, we present the development of a fluorescence-based directed evolution platform to evolve RTs that can both efficiently read-through m¹A and generate robust mutation signatures for NGS applications. We validated two of the best RT variants biochemically for their read-through properties and mutation signatures at m¹A sites. We deployed the best evolved RT in transcriptome-wide mapping of m¹A in HEK293T-derived mRNA at single-base resolution by NGS, denoted as “m¹A-IP-seq”. We confirmed many of the previously reported sites, and discovered hundreds of new m¹A sites in human mRNA. Finally, we developed “m¹A-quant-seq” that allowed us to estimate m¹A stoichiometries at individual sites in the transcriptome.

RESULTS

Choice of RT to Evolve

Retrovirus RTs and their engineered variants, are most commonly used in NGS library preparations. Among these RTs, human immunodeficiency virus (HIV) RT showed great potential in read-through efficiency over bulky modifications, such as 2' adducts²³. Although the native HIV-1 RT functions as a heterodimer containing two subunits, a 66 kDa catalytic domain (p66) and a 51 kDa structurally stabilizing domain (p51)^{24–26}, it was shown to be active as a homodimer with only the catalytic subunit²⁶. Additionally, high expression level of p66 in *E. coli* facilitates screening efforts. Therefore, we chose to evolve the p66 subunit of HIV-1 RT.

Fluorescence-based Assay for Mutation Detection

To carry out the directed evolution of RTs, we developed a fluorescence-based assay that can directly detect RT mutations. We deployed the Broccoli RNA fluorogenic aptamer, which becomes fluorescent upon binding to the fluorophore DFHBI-1T²⁷. To detect modifications on A, we first found a U-to-A mutation at U15 in Broccoli sequence that completely abolishes the fluorescence (Fig. 1a, Supplementary Fig. 1 and Supplementary Notes). When an RT generates mutations at m¹A15 reverting m¹A15 to U15, RNA will recover fluorescence in the presence of DFHBI-1T.

To detect such RT mutations, we designed a “RT-PCR-IVT” assay, where RT of m¹A RNA is performed followed by a polymerase chain reaction (PCR) and *in vitro* transcription (IVT) (Fig. 1b). First, the modified RNA is converted to the cDNA by an RT variant, which may generate truncated cDNA product in presence of m¹A (Fig. 1b). Mutations generated by RT will be encoded in the cDNA sequence. Next, only the read-through cDNA is amplified during PCR. PCR products contain the full-length Broccoli sequence with a 5' T7 promoter and maintain mutations generated during the previous RT step (Fig. 1b). Finally, the double-stranded DNA is transcribed into RNA using T7 RNA polymerase in the presence of DFHBI-1T. The presence of A-to-T mutations can be detected by fluorescence (Fig. 1b). Notably, the number of PCR cycles (PCN) is tunable, which controls the extent to which the cDNA product is amplified and is used as an adjustable selection pressure during the evolution.

To test the feasibility of the RT-PCR-IVT assay, we performed reactions using U15 and A15 RNA (sequences in Supplementary Table 1) as positive and negative RNA controls for 100% and 0% A-to-T mutation respectively with wild-type HIV RT (Fig. 1c). The assay has a large dynamic range and low fluorescence background (Fig. 1c). We performed the RT-PCR-IVT assay without intermediate purification steps and products of each step were validated using electrophoresis with controls (Fig. 1d and Supplementary Notes). Performing the assay with the m¹A15 substrate and wild-type HIV RT resulted in negligible fluorescence predominantly due to RT truncations (Fig. 1c, d and Supplementary Fig. 2a).

Directed Evolution of HIV-1 RT

We validated that the RT-PCR-IVT assay can work robustly using enzymes expressed in crude cell lysates, which further facilitated rapid screening (Supplementary Notes and Supplementary Fig. 2b). To control for any technical variations that may arise during multi-step reactions, we designed six technical controls on each 96-well culture plate with positive and negative control RNA, and wild-type HIV RT (Supplementary Fig. 2c). Note that we only compared fluorescence intensities within each plate unless otherwise specified.

We initiated our directed evolution efforts with eight libraries of HIV RTs; each library contains two randomized amino acid sites selected based on proximity to the replication base pair (bp) in the crystal structure of HIV RT (PDBID: 1RTD, Fig. 2a and Supplementary Table 2). The directed evolution was carried out by culturing the RT variants in 96-deep well plates. After growth, cell lysates were harvested on plate to perform subsequent RT and PCR reactions in test tubes, and finally T7 reactions on a 384-well plate. Fluorescence signal was recorded in real time on plate reader (Fig. 2b and Online Methods). In the first round of selection, Libraries 2 and 5, encompassing mutations at D76/R78 and W229/M230, respectively, showed positive and reproducible fluorescence responses (Fig. 2c and Supplementary Fig. 2d). In the second (Libraries 9–11) and third (Libraries 12 and 13) rounds of evolution, we saturated the key sites to find optimal mutations focusing on positions 76/78 and 229/230, respectively. Results suggested that the most promising variant contained the mutation pattern D76A, R78K, W229Y, and M230L (Supplementary Notes). In the third round (Libraries 14–17), we also screened sites L74, V75, F77, E79 and L80 surrounding 76 and 78 (Supplementary Table 2). RT-973 (V75F/F77A) yielded further improvements in fluorescence signal (Fig. 2c).

Previous studies of polymerase engineering and structural characterizations suggest mutations altering the interactions between domains around the active site of the polymerase may affect the template selectivity^{28,29}. In our case, the six favorable amino acid mutation sites located on both sides of the replication base pair (PDBID: 1RTD) and may contribute to decreased geometric constraints at the active site collectively, thereby accommodating m¹A during replication. We ultimately combined all six mutations that showed improved fluorescence in three rounds of selection, which yielded variant RT-1306 containing six mutations in total: D76A, R78K, W229Y, M230L, V75F, and F77A (Fig. 2c and Supplementary Table 1). We purified RT-1306, as well as RT-733 as one representative variant along the selection, for further biochemical characterization.

Biochemical Characterization of Evolved RTs

The RT-PCR-IVT assay using both purified enzymes (Supplementary Fig. 3a) showed over 10^4 -fold higher fluorescence response over the wild-type HIV-RT p66 at PCN = 4 (Fig. 3a, Supplementary Fig. 3b and Supplementary Notes), consistent with what we observed in the lysate assay during the screen. Sanger sequencing of the RT-PCR product of m¹A15 RNA suggested A-to-T mutation signature at the m¹A site for RT-1306 and RT-733, rather than the wild-type RT (Fig. 3b, Supplementary Fig. 3c and Supplementary Notes). We evaluated the read-through efficiency of RT by electrophoresis of the cDNA and RT-1306 showed a notably higher read-through efficiency comparing to RT-733 and the wild-type (Fig. 3c).

To rule out the possibility that the read-through and mutation properties of the evolved RTs are specific to the sequence of m¹A15 we used in the screen, we used a 45-mer synthetic RNA oligonucleotide library, different in sequence from m¹A15, each containing a single m¹A (“ModSig-m¹A”) flanked by 2-nucleotide random sequences (“-NNm¹ANN-”), where N = A/G/C/U (Supplementary Table 1). The control library “ModSig-A” contains same sequences but with unmodified A residues (“-NNANN-”). The read-through assay data from these libraries (Fig. 3d) showed that the wild-type RT produces predominantly truncated cDNA product, while RT-1306 consistently shows a higher read-through efficiency (~80%) than that of RT-733 (~40%).

To further assess mutation rates ($Mutation\ Rate = (countT + countC + countG) / (total\ count)$) at the m¹A in varying sequence contexts, we performed high-throughput sequencing of cDNA libraries generated from the synthetic ModSig RNA libraries using wild-type RT and the two evolved variants. The sequencing results revealed a moderate increase in overall mutation rate at the m¹A site from 66% in the wild-type RT library to ~84% in libraries built with the evolved variants (Fig. 3e and Supplementary Fig. 3d). The evolved RT variants showed a significantly enhanced A-to-T mutation rate (36% for wild-type versus 60% for RT-733 and RT-1306), which is the mutation signature that favors the fluorescence signal during the directed evolution screening. Most importantly, this elevated mutation rate is consistent across the 256 different sequence contexts, with no strong bias for the GGm¹ACC Broccoli sequence used in the evolution (Fig. 3e and Supplementary Fig. 3d, marked by a box).

In summary, in addition to significantly improved read-through efficiency, the mutational signatures of the evolved RT variants are significantly increased, without substantially affecting the background mutation rate (Supplementary Fig. 4). RT-733 and RT-1306 present comparable overall mutation rates based on the sequencing data, while RT-1306 possesses a higher read-through efficiency. *In vitro* comparisons of RT-1306 and TGIRT revealed that RT-1306 yielded 10-fold higher full-length cDNA product and a substantially higher ratio of read-through to truncation product compared with TGIRT (Fig. 3f and Supplementary Notes). Together, these data suggest RT-1306 may be capable of detecting m¹A in biological RNA samples.

Mutation Signatures in m¹A-IP-seq and m¹A-quant-seq

We applied RT-1306 in NGS libraries for m¹A mapping in mRNA. For m¹A-IP-seq, IP for immunoprecipitation with an anti-m¹A antibody, we used a protocol similar to the previously-described “m¹A-MAP” protocol¹¹ with optimized library construction steps for RT-1306. We treated RNA with the bacterial m¹A demethylase, α -ketoglutarate-dependent dioxygenase (AlkB) as a control for identifying m¹A-specific mutation signatures (Supplementary Fig. 5 and Supplementary Notes). For m¹A-quant-seq, we did not perform m¹A-IP; rather, we spiked in synthetic m¹A oligonucleotides with various m¹A fractions (i.e. “spike-in RNA”) to use for estimating m¹A stoichiometry (Fig. 4a and Online Methods). m¹A-IP-seq and m¹A-quant-seq protocols yielded high-quality libraries with expected library length (~250 bp), high alignment rate to the genome (Supplementary Table 3), wide transcriptome coverage, and good reproducibility of expressed transcripts between cell culture replicates (Supplementary Fig. 6 and Supplementary Notes).

To benchmark m¹A mutation signatures, we first manually examined the well-characterized m¹A sites in rRNA and tRNA. The mutation rate at m¹A1322 in 28S rRNA is 78% and 67%, which decrease to 21% and 5% after AlkB treatment, measured by m¹A-IP-seq and m¹A-quant-seq respectively (Fig. 4b). Interestingly, the mutation signatures at m¹A showed predominantly A-to-T mutation (Fig. 4b), different from major A-to-G mutation observed at the same site by the TGIRT^{11,21}. We observed robust mutation signatures ($75 \pm 11\%$; mean \pm s.d.) for m¹A58 across 38 cytosolic tRNA sequences (Fig. 4c), all of which also show a robust (~86%) reduction in mutation rates upon AlkB treatment (Fig. 4c). Mutation signatures and sensitivity to AlkB treatment are robustly observed with the spike-in m¹A RNA (Fig. 4d).

To assess the basal mutation level generated by the evolved RT, we examined mutation rates of all A residues in the spike-in RNA samples. The mean background mutation rate was found to be $0.25 \pm 0.16\%$ according to the synthetic spike-in samples (Fig. 4d). Similar background mutation levels were also observed for A sites in the context of rRNA ($0.4 \pm 0.01\%$), tRNA ($0.1 \pm 0.4\%$), and GAPDH mRNA ($0.6 \pm 0.01\%$). Taken together, these reference sites showed that the evolved RT-1306 can generate robust and accurate mutation signatures at m¹A sites in biological RNA samples with low background mutation.

Mutations and Read-through of m¹A Sites in mRNA and lncRNA

We examined mutation signatures in our sequencing datasets for three commonly found m¹A sites in prior single-base m¹A mapping data: PRUNE (annotated as both mRNA and ncRNA in the RefSeq database); MALAT1 (lncRNA); and ND5 (mitochondrial mRNA). For these examined sites, AlkB treatment shows 50+% decrease in mutation rate in both m¹A-IP-seq and m¹A-quant-seq ($p < 0.01$, one-sided t -test shown in Fig. 4e). m¹A-IP significantly enriches mutation rates for sites in PRUNE and ND5, but not in MALAT1 (Fig. 4e). This is not unexpected because m¹A-IP can enrich the m¹A fraction for sites starting with low m¹A stoichiometry, leading to increased mutation rates in IP libraries comparing to non-IP. Sites with high m¹A stoichiometry, such as MALAT1 case, may generate maximal mutation rates that will not further increase with IP. Importantly, visualization of mutations corresponding to these sites in the Integrative Genomics Viewer (IGV) revealed that they do

not fall on the ends of reads, suggesting that the mutation signatures mostly originate from read-through cDNA products (Supplementary Fig. 7).

Statistically Significant m¹A Sites from m¹A-IP-seq

Since m¹A was found as a relatively rare modification in mRNA and lncRNA (~0.014±0.004% m¹A/A in polyA-enriched RNA from HEK293T cells measured by LC-MS/MS, Supplementary Fig. 5a), we used m¹A-IP-seq to identify m¹A sites with increased sensitivity. We validated the anti-m¹A antibody activity *in vitro* by dot blotting and LC-MS/MS quantification following immunoprecipitation (Supplementary Fig. 8 and Supplementary Notes). With m¹A-IP-seq data, we confirmed the enrichment activity of the antibody by manual inspection and Lorenz curves analysis³⁰ (Supplementary Fig. 8). Although these tests suggest m¹A antibody can enrich m¹A-containing RNA fragments at examined sites, it remains challenging to evaluate its specificity for m¹A in the complex context of the human transcriptome. To identify m¹A sites, we relied on the mutation rate (1%) and its sensitivity to AlkB treatment evaluated by beta-binomial regression test over three replicates (Supplementary Fig. 9 and Online Methods). The mutation signatures will help rule out most interfering signals from non-specific antibody binding behavior.

In total, we identified 548 statistically significant m¹A sites ($p < 0.05$; Fig. 4f) from m¹A-IP-seq; these include 215 mRNA and 66 ncRNA sites (50 annotated as lncRNA in the NONCODE v5 database³¹), 194 cytosolic tRNA sites, and 57 mitochondrial sites, 16 intronic as well as 17 intergenic sites (Fig. 4f and Supplementary Table 4). Mutation signatures of sites are robustly captured by RT-1306 in comparison with dataset obtained by TGIRT; however, data overlap decreases mostly due to variations caused by AlkB treatment and antibody enrichment efficiencies (Supplementary Notes). We confirmed that these mutation sites are in the middle of aligned reads, but not the ends, again indicating that signals arise from mostly read-through cDNA products, rather than truncations (Supplementary Fig. 7b). Motif enrichment analysis of 281 m¹A sites in mRNA and ncRNA identifies a GA-rich motif (Supplementary Fig. 10a), similar to that identified in the m¹A-ID-seq⁸. 215 mRNA sites were found in 5'-UTR, 3'-UTR and CDS, with the largest number of sites found in CDS. The majority of CDS sites present low mutation rates (< 3%) suggesting low m¹A stoichiometry (Supplementary Fig. 10b).

Estimation of m¹A Stoichiometry Using m¹A-quant-seq

We applied the same statistical test in processing m¹A-quant-seq data. m¹A-quant-seq identified 55 potential mRNA and lncRNA m¹A sites in the HEK293T transcriptome with $p < 0.05$ (Supplementary Fig. 11a and Supplementary Table 5). We obtained a calibration curve using an approximated model correlating observed mutation rates and m¹A fraction to fit the values measured from the spike-in sample (Fig. 5a). We show estimation of m¹A fraction for a subset of 10 sites with relative high mutation rate and small p value identified by m¹A-quant-seq based on the calibration curve (Fig. 5b). The list of genes likely possessing m¹A at high and low fractions without antibody enrichment can serve as a useful reference list for future studies of m¹A biology. Encouragingly, mutation signatures and AlkB sensitivity observed by m¹A-quant-seq are highly comparable to those captured by a targeted library approach with deep coverage for PRUNE and MALAT1 m¹A sites

(Supplementary Fig. 11b,c), suggesting RT-1306 is applicable in locus-specific m¹A detection. We believe this method can be used to quantify and compare m¹A levels in different biological contexts, such as stress or disease.

DISCUSSION

Our data suggest that m¹A could be present in a large number of sites in human mRNA, although they might occur with low stoichiometry, especially for those we could only detect with anti-m¹A antibody enrichment. There could be new m¹A-specific mRNA methyltransferases yet to be discovered, or alternatively, m¹A sites could be installed by moonlight activities of tRNA modification enzymes. Previous studies have already revealed preferential m¹A modification in mitochondrial mRNA^{11,21}. Context-dependent RNA modifications is an emerging new theme in transcriptome regulation³².

Biological RNAs vary widely in endogenous abundance, chemical modifications, and in secondary and tertiary conformations. The diversity of the transcriptome encodes rich biological information, but imposes challenges in extracting precise RNA modification maps especially for less abundant modifications such as m¹A. By comparing our new datasets with previously obtained sequencing data from the same cell line, we found that although mutation signatures from mRNA can be robustly captured comparing RT-1306 with TGIRT, data reproducibility become worse when considering both immunoprecipitation and demethylation treatment steps^{33,34} (Supplementary Fig. 12 and Supplementary Notes). Our new antibody-free m¹A-quant-seq reduced such variance and enabled us to more directly estimate single-site m¹A stoichiometry.

Although mutation-based sequencing approaches have advanced the robustness in mapping of potential modified sites, there are potential concerns associated with peak calling as summarized by Sas-Chen *et al.*⁵. In our analysis pipeline, we used sequencing alignment with the soft-clipping option to minimize errors from non-templated addition (Supplementary Fig. 7), and deployed extensive annotation with various databases to decrease noise coming from SNPs and mis-annotations. Additionally, the beta binomial regression test on AlkB treatment sensitivity allows us to evaluate relative confidence level for each reported site. We do not rule out that detected sites can be mis-identified due to error-prone sequence contexts such as homopolymeric sequences, or ambiguous assignments coming from repeated sequences in the transcriptome such as MTRNR2 like genes. For such incidents, cross validations by orthogonal experiments or long-read sequencing would be required to confirm the presence and fraction of m¹A.

Noting that our m¹A modification fraction detection limit is ~8% (corresponding to 1% mutation rate), we likely underestimated the number of m¹A sites with low modification fraction. The current calibration curve fits a non-linear equation, which suggests that RT-1306 may still produce some degree of truncations in biological RNA contexts that decreases sensitivity at certain sites. Additionally, relying on demethylation treatment could lead to false negatives, especially for sites that are less abundant or located in complex structural contexts insensitive to enzymatic demethylation treatment. Future development of combined enzymatic and chemical demethylation can help improve sensitivity to these sites.

Finally, the current data processing requires sequence alignment with soft-clipping to avoid potential errors from non-templated addition by the RT; however, this will disable the detection of potential sites at or near the 5'-cap¹¹.

The evolved RTs will also find general applications beyond m¹A-seq, for example, in “DMS-seq” for RNA secondary structure probing^{36,37} where m¹A is the reaction product of more single-stranded A residues in RNA reacted with dimethyl sulfide (DMS). Finally, the selection method described here should be immediately deployable to evolve RTs that are sensitive to other types of RNA modifications. For modifications on bases other than A, appropriate mutations on the Broccoli aptamer may need to be identified first for modulating the fluorescence signal appropriately for the screen. Additionally, for modifications that do not perturb base pairing as much as m¹A, we anticipate larger library sizes will be needed to uncover sensitive enzymes. However, the robustness of the screening approach should make it amenable to automation, allowing us to screen much larger RT library sizes, which we are currently pursuing.

ONLINE METHODS

Sample Preparations

DNA and RNA Oligonucleotides—DNA primers, primers with FAM labeling and U15 RNA for *in vitro* assays and cloning were ordered from Integrated DNA Technologies, Inc (IDT) with standard desalting. Ligation adaptors used in m¹A-IP-seq and m¹A-IP-seq were ordered from IDT with HPLC purification. Other RNA oligonucleotides used in this study were synthesized in-house using an Expedite DNA synthesizer followed by normal deprotection for regular oligonucleotides and vendor-suggested deprotection for RNA oligonucleotides containing m¹A modifications to avoid Dimroth rearrangement. After deprotection, the RNA oligonucleotides were purified by HPLC with a C18 column and eluted with 0–20% acetonitrile in 0.1 M triethylammonium acetate. The desired peak was collected and dried by lyophilization. Synthesized RNA was dissolved in 10 mM Tris-HCl pH = 7.5, and the quality was examined by 10% 8M urea polyacrylamide gel electrophoresis (PAGE) gel. 33mer A15, m¹A15, and m¹A18 showed decent purity and were used directly. 43-mer RNAs in the spike-in samples showed significant impurity bands so we performed gel purification for all these RNAs with 10% urea PAGE gel and RNA recovery with the ZR small-RNA PAGE Recovery kit (Zymo Research).

Protein Expression and Purification—The 66 kDa subunit of the HIV-1 RT was cloned into a pET30a vector backbone with an additional 6xHis-tag on the N-terminus connected by a GGSG linker. The HIV-1 RT and its evolved variants were overexpressed in *E. coli* BL21 (DE3) cells and purified following the protocol reported previously³⁸. Briefly, two liters of cell culture was grown at 37°C for 3 hours in LB medium with 80 μM kanamycin until the OD₆₀₀ reached 0.5–0.6. Overexpression was induced by 0.5 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) and cells were incubated at 37°C for 2 hours and then at 16 °C overnight. Cells were harvested and resuspended in 40 mL lysis buffer (50 mM Na₂HPO₄ and NaH₂PO₄, 0.5 M NaCl, pH 7.8 dissolved half tablet of the proteinase inhibitor cocktail, Pierce) per liter of culture. The cells were then lysed by sonication and

centrifuged at 12,000 rpm for 40 minutes at 4 °C. Solubilized proteins in the supernatant were first purified using His60 Ni Superflow Resins (Clontech Laboratories, Inc) and eluted with 50 mM to 0.5 M gradient imidazole buffer containing 50 mM Na₂HPO₄ and NaH₂PO₄ pH = 6.0, 0.3 M NaCl and 10% Glycerol. The eluted protein fractions were run through a desalting column (PD-10, GE Healthcare), the buffer was exchanged into 3 mL ion-exchange Buffer A (50 mM Bis-tris pH 7.0) with additional 50 mM NaCl. Then, the fractions were subjected to Mono Q ion-exchange chromatography where the protein was injected onto the column flushing with 97.5% Buffer A and 2.5% Buffer B (50 mM Bis-tris pH 7.0, 1M NaCl) and the protein was recovered in the flow-through portion. The ion-exchange purification was found to be essential for effectively removing nuclease contamination. All purification steps were carried out at 4 °C or on ice. Fractions containing the expressed protein were combined and concentrated to 2.5 mL with a 30 kDa cut-off centrifugal filter (Millipore), run through the desalting column again, and was eluted with the storage buffer (50 mM Tris-HCl, 25 mM NaCl, 1 mM EDTA, 50% glycerol pH 7.0). Purified proteins were concentrated to 200 – 300 µL using a 30 kDa cut-off centrifugal filter (Millipore), aliquoted, flash frozen in liquid nitrogen and stored at –80 °C. Two liters of cell culture yielded around 3 mg of purified proteins for the wild-type HIV-1 p66 and the two evolved variants.

Wild-type T7 RNA polymerase with a 6xHis-tag attached at the N-terminus was cloned into the pBb vector³⁹. Protein over-expression in *E. coli* BL21 (DE3) cells was induced by 0.5 mM IPTG and 5 mM L-Arabinose at 30 °C overnight. The cells were lysed in lysis buffer (50mM Tris, pH 7.5, 1 M NaCl, 10 mM TCEP, 20% glycerol) by sonication, in the presence of a protease inhibitor cocktail (Pierce) and further purified with His60 Ni Superflow Resins. Proteins were eluted with 50 mM to 0.5 M imidazole gradient in the context of the lysis buffer followed by the desalting column and finally kept in a storage buffer composed of 50mM Tris-HCl pH 7.0, 100 mM NaCl, 1 mM EDTA, 10 mM DTT, and 50% glycerol at –80 °C.

AlkB demethylase was overexpressed and purified from *E. coli* BL21 (DE3) cells using an expression vector from Addgene (Plasmid #79050) following previously reported protocols^{9,11,40}. In short, overexpression was induced by adding 1 mM IPTG and 5 µM FeSO₄ into the cell culture, and cells were harvested 4 hours after induction. Cells were lysed by sonication in the lysis buffer (10 mM Tris-HCl pH 7.5, 300 mM NaCl, 5% glycerol, 2 mM CaCl₂, 10 mM MgCl₂, 10 mM 2-mercaptoethanol). Similar to the RTs, the purification of AlkB went through the Ni resin-desalting and then ion-exchange-desalting procedure. Note that the His60 Ni Superflow Resins were used with Buffer A (10 mM Tris-HCl pH 7.5, 1 M NaCl, 5% glycerol, 10 mM 2-mercaptoethanol) and Buffer B (10 mM Tris-HCl pH 7.5, 1 M NaCl, 5% glycerol, 10 mM 2-mercaptoethanol, 500 mM imidazole) instead. The buffers for the ion-exchange include Buffer A (20mM Bis-tris pH 7.0, 10mM 2-mercaptoethanol) and Buffer B (20mM Bis-tris pH 7.0 10mM 2-mercaptoethanol 1M NaCl). The purified protein was flash frozen and stored in lysis buffer supplemented with 30% glycerol at –80 °C.

Directed Evolution

Buffers—1x RT reaction buffer is composed of 50 mM Tris-HCl (pH 8.3), 75 mM KCl and 2mM MgCl₂. 1x PCR buffer was diluted from the 5x HF buffer (ThermoFisher) supplemented with 1 mM MgCl₂. 1x *in vitro* transcription buffer contains 40 mM Tris-HCl (pH 7.9), 30 mM MgCl₂, 2 mM Spermidine, 20 mM DTT.

Cloning and Preparation of Crude Lysate for RT Variants—Libraries were constructed by the Gibson assembly method using primer pools that contain NNK (or MNN when on the anti-sense strand) at the targeted mutation sites. Single colonies of NNK libraries were picked into 96-deep well plates and grown overnight with shaking at 37 °C. Overnight cultures were diluted by 15-fold with LB medium with kanamycin antibiotic, and then grown for 2.5 hours at 37 °C. Cells from the rest of the overnight cell culture were harvested by spinning down the plate at 4,000 rpm for 20 minutes at 4 °C and kept at –80 °C for determining genotypes of potential hits. We added 0.5 mM IPTG to each well to induce the overexpression of RT variants for 4 hours. The cells were then harvested by spinning down the deep well plates at 4,000 rpm for 20 minutes at 15 °C and the supernatant was discarded by decanting the plate. The cells were kept at –80 °C and lysed immediately before the screening assays. Harvested cells were lysed on the deep well plate by adding 80 µL lysis buffer containing 50 mM sodium phosphate, 300 mM NaCl, proteinase inhibitor cocktail (Thermo Scientific), and 1mg/mL lysozyme (Fisher Scientific), into each well. Cells were resuspended by vortexing the plate until no visible cell pellets were present at the bottom of the plate. The plate was subsequently incubated in the shaker at 37 °C for 1 hour, and then spun down at 4,000 rpm for 80 minutes at 4 °C. We then pipetted out 40 µL cell lysate from the supernatant for the following screening assays.

Screening Assays—The reverse transcription assays with the crude lysate screen were performed in 10 µL volumes containing 0.5 µM RT primer and RNA substrate, 4 mM deoxyribonucleoside triphosphate (dNTP) mixture (1 mM each dNTP), and 3 µL crude lysate in 1x RT reaction buffer. RT reactions were performed at 37 °C for 40 min, followed by denaturation at 80 °C for 10 min. Next, 1 µL volumes from the RT reactions were subjected to PCR amplification in 10 µL total volume, containing 1x PCR reaction buffer, and 4 µM each forward and reverse PCR primers, with a given number of PCR cycles. PCR reactions were carried out with a 30-second annealing step at 58 °C and 45 second elongation step at 72 °C. Finally, 7 µL of each PCR reaction mix was added as template into a 20 µL *in vitro* transcription reaction on a 384-well plate with glass bottom (Cellvis), together with 8 mM ribonucleoside triphosphate (rNTP) mixture (2 mM each rNTP), 50 µM DFHBI-1T (LuceRNA, Inc.) and 0.5 µL purified T7 RNA polymerase. The *in vitro* transcription reactions were monitored by plate reader (BioTek, Inc.) for 1.5 – 3 hours, one read per minute interval, with the excitation and emission wavelengths at 472 nm and 507 nm²⁷, respectively.

Biochemical Assays

Read-through Assay—The RT read-through assays were performed in 10 µL reaction volumes containing 1x RT buffer, 4 pmol RNA substrates and 2 pmol DNA primer with 5'-fluoresceindT label, 1 mM of each dNTP, and 2 µM purified RT. The RNA substrate and

DNA primer were added first and heated to 65 °C for 5 minutes and then annealed on ice for another 5 min. The RT reactions were then carried out at 37 °C for 1 hour followed by heating at 80 °C for 10 minutes to denature the RT. 1.5 µL of Proteinase K (Thermo Fisher) was added to the reaction mixture and then further incubated at 37 °C for 30 minutes for protein digestion. 4 µL of the reaction was mixed with 4 µL 2x RNA loading dye and heated to 95 °C for 5 min. 7 µL of that mixture was then immediately loaded onto 10% denaturing PAGE gel and run at 200V for 50 minutes. The gel was then imaged on a Bio-Rad imager.

AlkB treatment—AlkB treatment of RNA *in vitro* was performed in 40 µL reactions containing 10 – 50 pmol RNA substrate depending on the detection assay, purified AlkB (at least 4-fold of the RNA molarity), and 0.5 µL SuperaseIn RNase Inhibitor in 1x reaction buffer. For initial tests of purified AlkB activity, 1x reaction buffer contained 300 mM KCl, 2 mM MgCl₂, 50 µg/mL BSA, 50 µM (NH₄)₂Fe(SO₄)₂, 50 mM MES pH 5.0, 300 µM 2-ketoglutarate, and 2 mM L-ascorbic acid. During later optimization assays, reaction buffer components were altered as described in Supplementary Fig. 5. The AlkB treatment reactions were carried out at 25 °C for 2 hours unless otherwise specified, and then quenched by adding 4 µL 50 mM EDTA. The RNA was purified by an Oligo Clean and Concentrator kit (OCC, Zymo Research), following the kit instructions and eluted with 15–30 µL RNase-free water for the following assays.

LC-MS/MS—50–200 ng of RNA were first digested by nuclease P1 (Wako, 1U) in 26 µL 1x P1 digestion buffer containing 25 mM NaCl, 2.5 mM ZnCl₂ at 42 °C for 2 hours. Next, 1 µL FastAP Thermosensitive Alkaline Phosphatase (ThermoFisher, 1 U/µL) and 3 µL 10x FastAP buffer (ThermoFisher) were added to the reaction and incubated at 37 °C for 2 hours. Reactions were then diluted into 60 µL and filtered by a 0.22 µm filter (Millipore). Filtered samples were injected into a C18 reverse phase column (Agilent) on a UHPLC coupled to an Agilent 6460 or a SCIEX 6500+ Triple Quad Mass Spectrometer in positive electrospray ionization mode. Quantitation was performed based on peak areas of characteristic nucleoside-to-base transitions, including 268-to-136 for A, 284-to-152 for G, 282-to-150 for m¹A and m⁶A with retention time at 0.8 minutes and 2.8 minutes, respectively.

Dot blotting—5mer synthetic RNA oligonucleotides were made into gradient concentrations in 10, 8, 4, 2, 1, 0.5 ng/µL and 1 µL of each was dotted on a Hybond-N membrane (GE Healthcare) optimized for nucleic acid transfer. Samples were left on the benchtop for around 5 minutes to dry completely and then UV-crosslinked to the membrane by running auto-crosslinking mode twice. The membrane was then blocked with 5% milk in 1x PBST for 1 hour at room temperature; washed with 1x PBST for 4 times with 10 minutes each time. The membrane was incubated with m¹A antibody 0.1 µg/mL in 3% BSA in 1x PBST at 4 °C overnight while shaking, then washed by 1x PBST 6 times, and incubated with the secondary anti-mouse IgG antibody in 1% milk and 1x PBST for 1 hour at room temperature. Membrane was washed with 1x PBST for 3 times and developed using 1:1 SuperSignal™ West Pico PLUS substrate and peroxide solution, then imaged by the FluorChem R imager under chemiluminescence detection.

m¹A Immunoprecipitation for LC-MS/MS—Total RNA extracted HEK293T cells with Trizol reagent was first treated with DNase I and then small RNA fractions were removed using the MegaClear kit. Remaining total RNA were then fragmented using the magnesium fragmentation module into 100–150 sized fragments by heating at 94 °C for 5 minutes in 1x fragmentation buffer from NEBNext Magnesium RNA Fragmentation Module. Fragmentation was quenched by adding 1x stop buffer immediately to the RNA, and cleaned up with the OCC kit (Zymo Research). 40 µg of fragmented RNA was used as “Input” RNA for m¹A immunoprecipitation (IP). RNA was incubated with 10 µg of m¹A antibody (MBL 345–3) in 400 µL 1x IP buffer (10 mM Tris-HCl pH 7.4, 150 mM NaCl, and 0.1% NP-40) at 4 °C for overnight or 1 hour. After incubation, 10 µL protein G beads were washed, resuspended in 400 µL 1x IP buffer, and added to RNA and antibody to incubate at 4 °C for another 3 hours while shaking. Each IP was washed with 1 mL 1x IP buffer four times followed by 1 mL 1x TEN buffer (10 mM Tris-HCl pH 7.4, 1 mM EDTA, and 0.1% NP-40) twice. Elution was performed by incubating the beads with 5U proteinase K in 100 µL 5 mM Tris-HCl pH 7.4, 1 mM EDTA and 0.25% SDS at 37 °C for 1 hour while shaking. The elution mixture was further purified by the OCC kit (Zymo Research). All eluted RNA was then subjected to the LC-MS/MS assay.

Library Preparations for Next-generation Sequencing

Synthetic Oligonucleotide Library—100 ng of pooled synthetic RNA oligonucleotides containing both ModSig-m¹A and ModSig-A sequences (Supplementary Table 1) was subjected to the NEBNext Small RNA Library Prep Set for library construction for Illumina sequencing following the manufactural instruction for all steps with following changes. First, we used customized 3′ and 5′ ligation adaptors containing an additional NNNNN randomized sequence as well as a unique barcoding sequence (Supplementary Table 1) in order to alleviate sequence biases during ligations and facilitate PCR duplicate removal in the data processing. Secondly, we replaced the RT in the RT step with 1 µL 10 µM purified wild-type HIV-1 RT or evolved RT variant. The RT reaction was performed following both 3′ and 5′ adaptor ligations, directly followed by PCR amplification; therefore, read-through cDNA products rather than truncated cDNAs were amplified and sequenced.

Biological RNA Library—Step-by-step protocols for m¹A-IP-seq and m¹A-quant-seq are described in the Extended Protocols. For each biological replicate, 5 mg of total RNA were extracted from HEK293T cells using TRIzol (Invitrogen) reagent and the MaXtract High Density tubes (Qiagen) following manufactural instructions. We then performed DNase I (Thermo Fisher) digestion followed by an ethanol precipitation step to recover RNA (Extended Protocols in Supplementary Notes). Next, we performed small RNA removal to remove tRNA and small RNA species using the MEGAclear Transcription Clean-up kit (Invitrogen, AM1908) as previously described¹¹. Next, we carried out two rounds of polyA enrichment to purify mRNA using oligo (dT) DynaBeads (Invitrogen) with 3 mg total RNA after small RNA removal. The mRNA was then fragmented to ~100–150 nt by heating at 94 °C for 5 minutes using the 1x Magnesium RNA fragmentation module (NEB)¹¹, stopped with the 1x stopping solution in the module and cleaned up using the OCC (Zymo Research). RNA fragment yields at this step were around 47 – 60 µg and we took 30 µg of

the fragments for the immunoprecipitation (IP), keeping the rest for m¹A-quant-seq (Fig. 4a).

For m¹A-IP-seq, we carried out IP by mixing 30 µg polyA-enriched RNA with 10 µg m¹A antibody (MBL D345–3) in 1x IP buffer (10 mM Tris-HCl pH 7.4, 150 mM NaCl and 0.1% NP-40) with the SuperaseIn RNase inhibitor (Thermo Fisher Scientific) and incubating at 4 °C for 1 hour while rotating. Protein G magnetic beads were added afterwards and incubated together with RNA and antibody for three hours. The beads were washed with 1 mL 1x IP buffer four times and 1 mL TEN buffer (10 mM Tris-HCl pH 7.4, 1 mM EDTA pH 8.0 and 0.05% NP-40) twice, and finally eluted with 200 µL 1x elution buffer (1x IP buffer with added 10 mM N¹-methyladenosine) twice by rotating at room temperature each time. Eluted IP RNA was cleaned up with the OCC kit (Zymo Research) and finally eluted in 30 µL RNase-free water (Thermo Fisher Scientific). At this step, 10 µL of IP RNA was saved for the “-AlkB” library while 20 µL were subjected to AlkB treatment. The AlkB treatment was performed in the freshly prepared demethylation reaction buffer (50 mM MES pH 5.0, 283 µM (NH₄)₂Fe(SO₄)₂·6H₂O, 0.3 mM 2-ketoglutarate, 2 mM L-ascorbic acid, 10 mM KCl, 2 µL Superase In, ~40 pmol AlkB protein), incubating at 25 °C for 2 hours. AlkB reactions were quenched with 5 mM EDTA, cleaned up with the OCC kit (Zymo Research) and eluted with 16 µL RNase-free water as the “+AlkB” IP RNA. The IP RNA “-AlkB” and “+AlkB” samples were subjected library constructions, respectively, following the published m¹A-Map ligation-based protocol¹¹ with optimized purification procedures and the evolved RT-1306 (Supplementary Fig. 6a and Extended Protocols). For the m¹A-quant-seq, we added 1 ng of spike-in RNA into 100 ng of polyA RNA fragments in 30 µL for each biological replicate. Two thirds of RNA with spike-in was subjected to AlkB treatment and clean-up at conditions described in m¹A-IP-seq, and made into “+AlkB” library; the rest of RNA fragments with spike-in were used for “-AlkB” library. Final library sizes for all replicates ranged ~220 bp shown by the bioanalyzer data (Supplementary Fig. 6).

Targeted Library—RNA was purified as for m¹A-IP-seq until the fragmentation step. ~300 ng polyA RNA fragments were then subjected to AlkB treatment and purification. 150 ng polyA RNA fragments with and without AlkB treatment were treated with PNK for ends repair, ligated with 3′-adaptor, reverse transcribed into cDNA with RT-1306 and purified as in the m¹A-IP-seq library approach (Extended Protocols). Purified cDNA was PCR amplified with locus-specific PCR primers (Supplementary Table 1). For each targeted site, individual PCR was performed with Phusion DNA polymerase. PCR reactions were then combined and purified. 14 ng DNA products from both “-AlkB” and “+AlkB” RNA were built into libraries with the NEBNext Ultra II DNA library Prep kit (NEB 7645).

NGS Data Processing and Analysis

Synthetic RNA oligonucleotide libraries—The raw sequencing data were first processed by the “cutadapt” program⁴¹ with the option “-q 20 -m 30” to filter reads by the sequencing quality > 20 and minimum read length of 30. As both ligation adaptors contain a random 5-nt sequence at their termini, PCR duplicates can be identified if two reads are completely identical including the random sequences. Therefore, PCR duplicates were subsequently removed using the “fastx_collapser” program. The processed sequencing data

were further analyzed by the in-house python scripts “countmu.py” to count mutations at the m¹A sites for each library, and “contextstats.py” to process mutation rates in the 256 dinucleotide sequence contexts.

Biological RNA libraries—Although pair-end 100 bp sequencing was performed with the dataset, we focused on processing the R2 reads due to the uncertainty in the location of UMI in the R1 reads as previously noted in m¹A-MAP processing¹¹. The flow chart summarizing the pipeline is summarized in Supplementary Fig. 10. Raw data (R2 reads of the pair-end data) were first processed with the “clumpify.sh” program in the BMap package (BMap-Bushnell [B.-sourceforge.net/projects/bbmap/](https://sourceforge.net/projects/bbmap/)) with the option “dedupe subs=0” to remove PCR duplicates. Adaptors were trimmed by the “cutadapt” program⁴¹ while reads were filtered by quality and length with options “-a AGATCGGAAGAGCGTCGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT -q 20 -m 30”. Reads were aligned to the hg19 genome file downloaded from the UCSC database as well as the known splicing sites via the hisat2 aligner with the option “--trim5 11 ” removing the unique molecular identifier (UMI) sequence with the script “run_cut_alignment_R2_hg19.sh” noting that default “soft-clipping” option is applied with the alignment. We found that turning on this option was crucial to rule out non-templated addition of RT at the ends that can lead to high probability false positive mutation rates, although we do not rule out that having this option on would likely decrease our sensitivity in detecting mutations towards the ends of the transcripts or RT truncation products with remaining mutation signatures. The resulting bam file was then split into positive and negative bam files, sorted and indexed using “samtools”⁴² with script “posneg.sh”. Next, we counted reads and identified initial mutation sites using the following scripts from the positive and negative bam files respectively. Read coverage at each base level throughout the hg19 genome was counted by the “bam-readcount” tool with the filter “-b 20” to only count reads with a minimum quality of 20 at each nucleotide position and the default maximum counting depth of 8000 for each location in reference to the script “bam-readcount.sh”. Sites that harbor an “A” in the reference genome for the positive bam file or a “T” for the negative bam, show mutation rate higher than 1% and are covered by at least 5 reads were extracted from the bam-readcount output files of the “-AlkB” libraries with the script “parse_pos.py” or script “parse_neg.py”; corresponding sites were then found accordingly in the “+AlkB” libraries via the script “control_parse_pos.py” or “control_parse_neg.py”. Finally, we used “clean_untreat.py” to clean up the non-overlapping sites resulting from the “-AlkB” and corresponding “+AlkB” library. The resulting counted sites are output as .csv files and subjected to m¹A identification using R scripts.

Base-resolution m¹A identification in m¹A-IP-seq and m¹A-quant-seq—In general, the resulting csv files that contain the initially extracted mutation sites were formatted with a “Prepare.R” script for subsequent analysis, and processed through the “Callm1A-ip.R” or “Callm1A-input.R” script for m¹A site identification based on defined criteria, gene annotation and statistical tests.

m¹A sites are pre-selected by the faithfulness of the detected mutation rate and the sensitivity to the demethylation treatment of a given site according to the following criteria:

1) single sites that contain at least 5 mutation counts in “-AlkB” libraries in all three biological replicates¹¹; 2) sites that possess at least 1% mutation rates in “-AlkB” libraries; 3) sites that are sensitive to AlkB treatment in three replicates, i.e. $Mutation\ Rate(-AlkB) - Mutation\ Rate(+AlkB) > 0$. 2041 sites were selected under these criteria for m¹A-IP-seq, and then annotated by the “ChIPseeker”⁴³ and “GenomicFeatures”⁴⁴ packages in R. Single RNA sites were annotated following the priority list: 1) the Single Nucleotide Polymorphism collected for HEK293T cells⁴⁵; 2) tRNA (hg19) downloaded from the UCSC table browser (<https://genome.ucsc.edu/cgi-bin/hgTables>); 3) Mitochondrial RNA; 4) cytosolic RefSeq-annotated mRNA database in the “hg19_UCSC.gtf” downloaded from the Illumina iGenomes website (https://support.illumina.com/sequencing/sequencing_software/igenome.html); 5) long non-coding RNA from the NONCODE v5 database³¹; 6) other Ref-seq annotated non-coding RNA according to the “hg19_UCSC.gtf”; 7) intergenic RNA for the rest of the non-annotated sites. We also crosschecked the resulting sites with the RNA editing database “RADAR”⁴⁶ for sites that are potentially assigned as RNA editing rather than m¹A. In principle editing sites should not show significant sensitivity to AlkB treatment; for sites that have strong AlkB sensitivity, we do not exclude that they could be true m¹A sites that were mis-assigned as editing. The resulting 2041 sites were tested by the beta-binomial regression⁴⁷ and reported in Supplementary Table 4 with *p* values. m¹A-quant-seq data were processed following the same pipeline, except using a cut-off of 3 mutation count rather than 5 during the pre-selection (Supplementary Table 5).

rRNA and tRNA analysis—Raw reverse (R2) reads with depleted PCR duplicates and removed adaptors were aligned to the 28S rRNA (NR_003287.4.fa) or tRNA (hg19-tRNAs.fa) genomes downloaded from the RefSeq database. We then counted reads at each base with the “bam-readcount” program in reference to the corresponding genome.

Lorenz curve analysis for antibody enrichment—We assessed antibody enrichment efficiency using an in-house script which has similar rationale with the ChIP-Seq quality control tool CHANCE³⁰. Basically, each transcript was divided into 300 bins (100 for 5′ UTR, 100 for CDS, and 100 for 3′ UTR), and reads on each bin were counted by Bedtools⁴⁸. To eliminate the impact of different genes expression, we have normalized the reads counts in IP samples with input samples. The top 100 k transcriptomic bins were used to generate Lorenz Curve.

Enrichment peak calling by MeRIPtools—We performed IP enrichment peak calling based on enrichment of read count in m¹A-IP-seq over m¹A-quant-seq by an in-house R-package “MeRIPtools”, which test for enrichment using a binomial-distribution-based model. This procedure resulted in 4178 enriched peaks called in at least two out of the three replicates with FDR < 0.05, 2185 of which show at least 2-fold averaged enrichment ($FC_{avg} > 2$, where $FC = Count_{IP-seq} / Count_{quant-seq}$) across biological triplicates.

Spike-in analysis and calibration curve for m¹A stoichiometry estimation—Sequencing data were aligned to the spike-in RNA sequences after deduplication and adaptor cut. Reads were counted by bam-readcount. The observed mutation rates at different m¹A fraction levels showed a non-linear trend, which we reason could arise from potential

truncation during RT or other biases from the following ligation or PCR steps. We adopted a simplified model to describe the observed mutation rate that mainly arises from read-through products over m¹A. We used a fraction parameter f_A to account for the fraction of read-through cDNA that came from m¹A, and assumed an averaged mutation rate of 100% m¹A as R_0 . The observed mutation rate y (in percentage) and m¹A fraction x (in percentage) can be expressed by the following equation,

$$y = \frac{B(100 - x) + xf_A R_0}{100 - (1 - f_A)x} \times 100$$

where B is the background mutation rate (i.e. mutation rate at A); $100 - x$ represents fraction of A; and $1 - f_A$ accounts for fraction short cDNA that could come from RT truncations or potential ligation or PCR biases in the experimental steps that do not contribute to the overall coverage at the m¹A site. We fit the spike-in mutation rate and m¹A fraction using the above equation with the non-linear curve fitting by the Prism software. The best-fit values from data resulted in $R_0 = 0.52$ and $f_A = 0.22$, which gives the calibration equation (Fig. 5a) for estimation of m¹A stoichiometry in m¹A-quant-seq.

Motif enrichment and Metagene analyses—9mer sequences with 4 nucleotides before and after the identified m¹A sites in m¹A-IP-seq with $p < 0.05$ were retrieved as in the fasta format. Sequences were subjected to the online MEME Suite program⁴⁹ for “Motif Discovery” with the default setup. Metagene profile was plotted with the “Guitar” R package⁵⁰s

Statistical tests

Data are presented as mean with s.e.m or s.d. as noted in respective case. All values of n are provided and no data were excluded.

Reporting Summary

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

Data availability

Raw and processed m¹A-IP-seq and m¹A-quant-seq data are available at NCBI Gene Expression Omnibus, accession number GSE123365. The DNA sequence of RT-1306 is shown in Supplementary Table 1, and the plasmid for bacterial expression of RT-1306 is available on Addgene with the ID 131521. The data that support the findings of this study are available from the corresponding author upon request.

Code availability

Processing scripts for synthetic m¹A oligonucleotide library, m¹A-IP-seq and m¹A-quant-seq are available in the Supplementary Data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

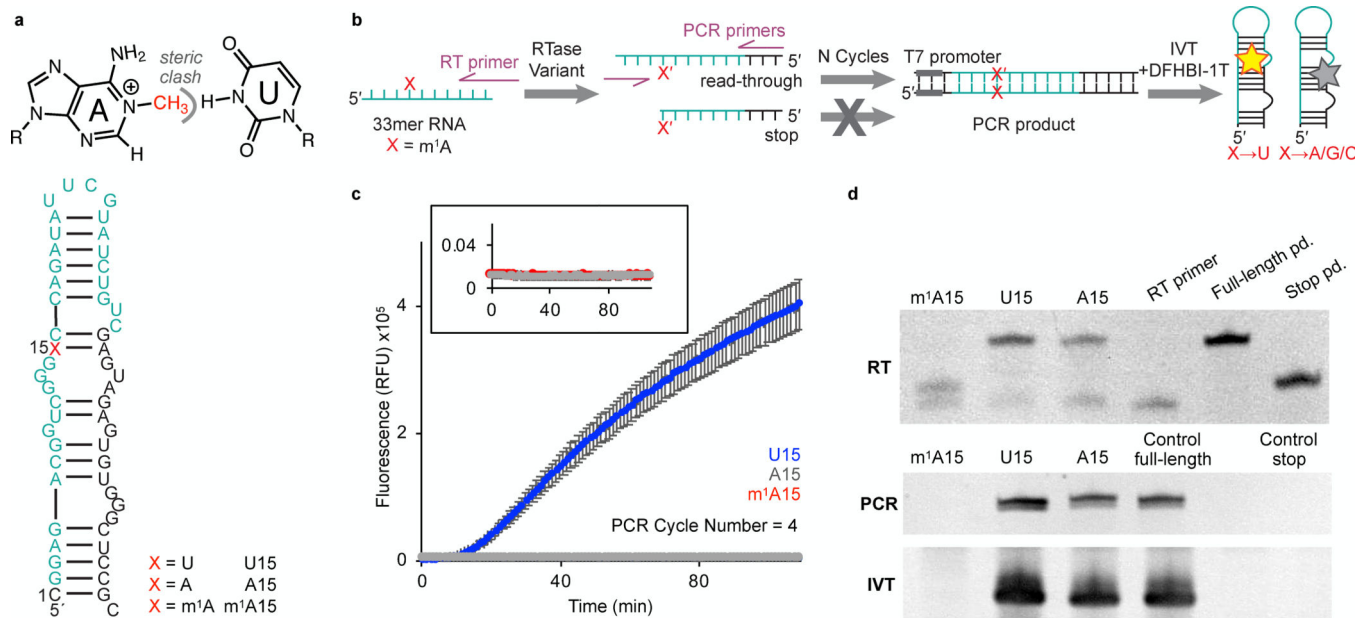
We would thank the entire staff at the UCCCC sequencing facility at the University of Chicago for performing the Sanger and NGS sequencing measurements; Professor Tao Pan for sharing the ModSig libraries; Dr. Lisheng Zhang and Chang Liu for sharing the protocols for constructing biological RNA libraries; Dr. Lulu Hu and Liangzhi Luo for helping with protein purification. This work was supported by the University of Chicago, National Human Genome Research Institute (RM1 HG008935, C.H. and B.C.D) and National Institute of General Medical Sciences (R35 GM119840, B.C.D.) of the National Institutes of Health, the University of Chicago Medicine Comprehensive Cancer Center (P30 CA14599), the Chicago Fellows Program (H.Z.) and the MSTP Training Grant (T32GM007281) (C. S.). S.N. is an HHMI fellow of the Damon Runyon Cancer Research Foundation (DRG-2215-15). C.H. is a Howard Hughes Medical Institute Investigator.

References

1. Roundtree IA, Evans ME, Pan T & He C Dynamic RNA Modifications in Gene Expression Regulation. *Cell* 169, 1187–1200, (2017). [PubMed: 28622506]
2. Frye M, Harada BT, Behm M & He C RNA modifications modulate gene expression during development. *Science* 361, 1346, (2018). [PubMed: 30262497]
3. Jonkhout N et al. The RNA modification landscape in human disease. *RNA* 23, 1754–1769, (2017). [PubMed: 28855326]
4. Dominissini D et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485, 201, (2012). [PubMed: 22575960]
5. Meyer KD et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* 149, 1635–1646, (2012). [PubMed: 22608085]
6. Dominissini D et al. The dynamic N1-methyladenosine methylome in eukaryotic messenger RNA. *Nature* 530, 441, (2016). [PubMed: 26863196]
7. Li X, Xiong X & Yi C Epitranscriptome sequencing technologies: decoding RNA modifications. *Nature Methods* 14, 23, (2016). [PubMed: 28032622]
8. Li X et al. Transcriptome-wide mapping reveals reversible and dynamic N1-methyladenosine methylome. *Nature Chemical Biology* 12, 311, (2016). [PubMed: 26863410]
9. Zheng G et al. Efficient and quantitative high-throughput tRNA sequencing. *Nature Methods* 12, 835, (2015). [PubMed: 26214130]
10. Cozen AE et al. ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nature Methods* 12, 879, (2015). [PubMed: 26237225]
11. Li X et al. Base-Resolution Mapping Reveals Distinct m1A Methylome in Nuclear- and Mitochondrial-Encoded Transcripts. *Molecular Cell* 68, 993–1005.e1009, (2017). [PubMed: 29107537]
12. Schwartz S & Motorin Y Next-generation sequencing technologies for detection of modified nucleotides in RNAs. *RNA Biology* 14, 1124–1137, (2017). [PubMed: 27791472]
13. Mohr S et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* 19, 958–970, (2013). [PubMed: 23697550]
14. Zhao C, Liu F & Pyle AM An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* 24, 183–195, (2018). [PubMed: 29109157]
15. Ellefson JW et al. Synthetic evolutionary origin of a proofreading reverse transcriptase. *Science* 352, 1590, (2016). [PubMed: 27339990]
16. Huber C, von Watzdorf J & Marx A 5-methylcytosine-sensitive variants of *Thermococcus kodakaraensis* DNA polymerase. *Nucleic Acids Research* 44, 9881–9890, (2016). [PubMed: 27651460]

17. Harcourt EM, Ehrenschwender T, Batista PJ, Chang HY & Kool ET Identification of a Selective Polymerase Enables Detection of N6-Methyladenosine in RNA. *Journal of the American Chemical Society* 135, 19079–19082, (2013). [PubMed: 24328136]
18. Helm M, Giegé R & Florentz C A Watson–Crick Base-Pair-Disrupting Methyl Group (m1A9) Is Sufficient for Cloverleaf Folding of Human Mitochondrial tRNALys *Biochemistry* 38, 13338–13346, (1999). [PubMed: 10529209]
19. Anderson JT & Droogmans L in *Fine-Tuning of RNA Functions by Modification and Editing* (ed Henri Grosjean) 121–139 (Springer Berlin Heidelberg, 2005).
20. Saikia M, Fu Y, Pavon-Eternod M, He C & Pan T Genome-wide analysis of N1-methyl-adenosine modification in human tRNAs. *RNA* 16, 1317–1327, (2010). [PubMed: 20484468]
21. Safra M et al. The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature* 551, 251, (2017). [PubMed: 29072297]
22. Xu L et al. Three distinct 3-methylcytidine (m3C) methyltransferases modify tRNA and mRNA in mice and humans. *Journal of Biological Chemistry*, (2017).
23. Patterson JT, Nickens DG & Burke DH HIV-1 Reverse Transcriptase Pausing at Bulky 2' Adducts is Relieved by Deletion of the RNase H Domain. *RNA Biology* 3, 163–169, (2006). [PubMed: 17396357]
24. Kohlstaedt LA, Wang J, Friedman JM, Rice PA & Steitz TA Crystal structure at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science* 256, 1783, (1992). [PubMed: 1377403]
25. Huang H, Chopra R, Verdine GL & Harrison SC Structure of a Covalently Trapped Catalytic Complex of HIV-1 Reverse Transcriptase: Implications for Drug Resistance. *Science* 282, 1669, (1998). [PubMed: 9831551]
26. Müller B et al. Co-expression of the subunits of the heterodimer of HIV-1 reverse transcriptase in *Escherichia coli*. *Journal of Biological Chemistry* 264, 13975–13978, (1989). [PubMed: 2474539]
27. Filonov GS, Moon JD, Svensen N & Jaffrey SR Broccoli: Rapid Selection of an RNA Mimic of Green Fluorescent Protein by Fluorescence-Based Selection and Directed Evolution. *Journal of the American Chemical Society* 136, 16299–16308, (2014). [PubMed: 25337688]
28. Chen T et al. Evolution of thermophilic DNA polymerases for the recognition and amplification of C2'-modified DNA. *Nature Chemistry* 8, 556, (2016).
29. Betz K et al. Structural Insights into DNA Replication without Hydrogen Bonds. *Journal of the American Chemical Society* 135, 18637–18643, (2013). [PubMed: 24283923]
30. Diaz A, Nellore A & Song JS CHANCE: comprehensive software for quality control and validation of ChIP-seq data. *Genome Biology* 13, R98, (2012). [PubMed: 23068444]
31. Fang S et al. NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Research* 46, D308–D314, (2017).
32. Shi H, Wei J & He C Where, When, and How: Context-Dependent Functions of RNA Methylation Writers, Readers, and Erasers. *Molecular Cell* 74, 640–650, (2019). [PubMed: 31100245]
33. Dominissini D & Rechavi G Loud and Clear Epitranscriptomic m1A Signals: Now in Single-Base Resolution. *Molecular Cell* 68, 825–826, (2017). [PubMed: 29220649]
34. Xiong X, Li X, Wang K & Yi C Perspectives on topology of the human m1A methylome at single nucleotide resolution. *RNA*, (2018).
35. Sas-Chen A & Schwartz S Misincorporation signatures for detecting modifications in mRNA: Not as simple as it sounds. *Methods* 156, 53–59, (2019). [PubMed: 30359724]
36. Zubradt M et al. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nature Methods* 14, 75, (2016). [PubMed: 27819661]
37. Guo JU & Bartel DP RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science* 353, (2016).
38. Le Grice SFJ, Cameron CE & Benkovic SJ in *Methods in Enzymology* Vol. 262 130–144 (Academic Press, 1995). [PubMed: 8594344]
39. Lee TS et al. BglBrick vectors and datasheets: A synthetic biology platform for gene expression. *Journal of Biological Engineering* 5, 12, (2011). [PubMed: 21933410]

40. Mishina Y, Chen LX & He C Preparation and Characterization of the Native Iron(II)-Containing DNA Repair AlkB Protein Directly from *Escherichia coli*. *Journal of the American Chemical Society* 126, 16930–16936, (2004). [PubMed: 15612731]
41. Martin M Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 17, 3, (2011).
42. Li H et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079, (2009). [PubMed: 19505943]
43. Yu G, Wang L-G & He Q-Y ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* 31, 2382–2383, (2015). [PubMed: 25765347]
44. Lawrence M et al. Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology* 9, e1003118, (2013). [PubMed: 23950696]
45. Lin Y-C et al. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nature Communications* 5, 4767, (2014).
46. Chen L Characterization and comparison of human nuclear and cytosolic editomes. *Proceedings of the National Academy of Sciences* 110, E2741, (2013).
47. Rigby RA & Stasinopoulos DM Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54, 507–554, (2005).
48. Quinlan AR & Hall IM BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842, (2010). [PubMed: 20110278]
49. Bailey TL et al. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research* 37, W202–W208, (2009). [PubMed: 19458158]
50. Cui X et al. Guitar: An R/Bioconductor Package for Gene Annotation Guided Transcriptomic Analysis of RNA-Related Genomic Features. *BioMed Research International* 2016, 8, (2016).

**Figure 1.**

Fluorescence-based RT mutation detection assay. **(a)** Chemical structure of m¹A and perturbation of canonical A-U base pair. The secondary structure of the Broccoli RNA aptamer, where the X position represents the site that modulates the fluorescence response upon mutation. The truncated 33-mer RNA substrate sequence used is colored in cyan. **(b)** Schematic of the RT-PCR-IVT assay. **(c)** Time-dependent fluorescence assay data of the RT-PCR-IVT assays with the positive (U15), negative (A15) and the screening substrate (m¹A15) RNAs. Data shown are fluorescence intensities in relative fluorescence units (RFU) with mean±s.e.m, from n = 5 independent assays. **(d)** Electrophoresis data for product characterization after each reaction step during the RT-PCR-IVT assay. The RT primer, full-length cDNA (“Full-length pd.”) and truncated cDNA (“Stop pd.”) oligonucleotides are used as references in the control lanes in the top gel; PCR products using the full-length (“Control full-length”) and truncated (“Control stop”) cDNAs as templates, as well as corresponding IVT RNA products, are shown in the control lanes in the middle and bottom gels respectively.

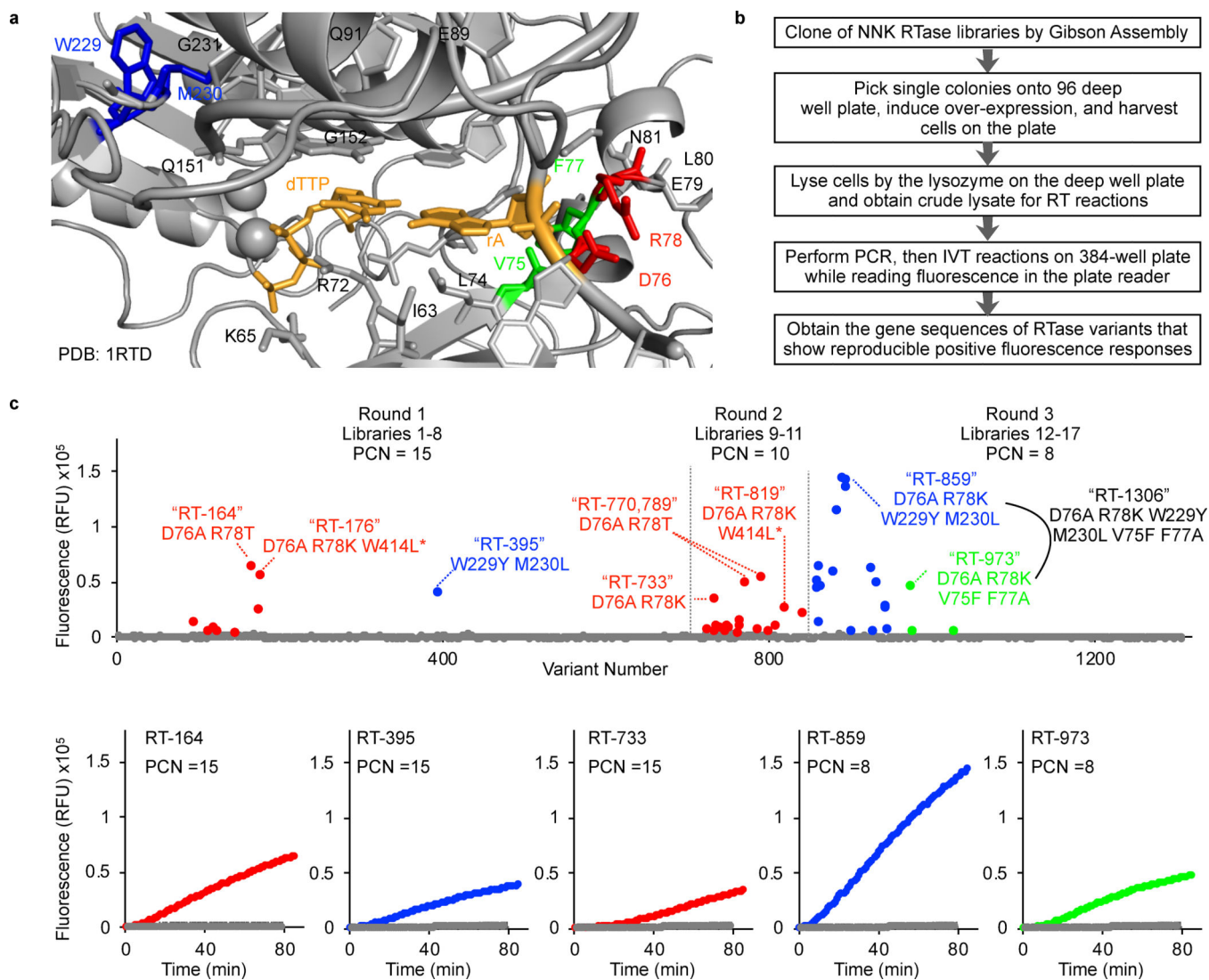


Figure 2. Design and results of the directed evolution of the HIV-1 RT. **(a)** Zoomed-in view of the active site from the crystal structure of the HIV-1 RT with a substrate bound (PDBID: 1RTD). The replicating base pair (template rA and incoming dTTP) is colored in yellow. Each RT library permute two amino acid sites and all amino acid sites screened are labeled with residue numbers (Supplementary Table 2). Mutation sites that show elevated fluorescence during screening in RT libraries 2, 5 and 14 are colored in red, blue, and green respectively. **(b)** Flow chart of the directed evolution procedure. **(c)** The upper panel presents the fluorescence intensity measured at 85 minutes during IVT for all the variants screened throughout the directed evolution. RT variants that show elevated fluorescence (at least 2-fold to wild-type RT) are shown in color and colors are coded by mutated amino acid sites in correspondence to those colored in panel a. The lower panel shows raw data curves during screening for representative RT variants; 2 more biological replicates of selected variants were shown in Supplementary Fig. 2e.

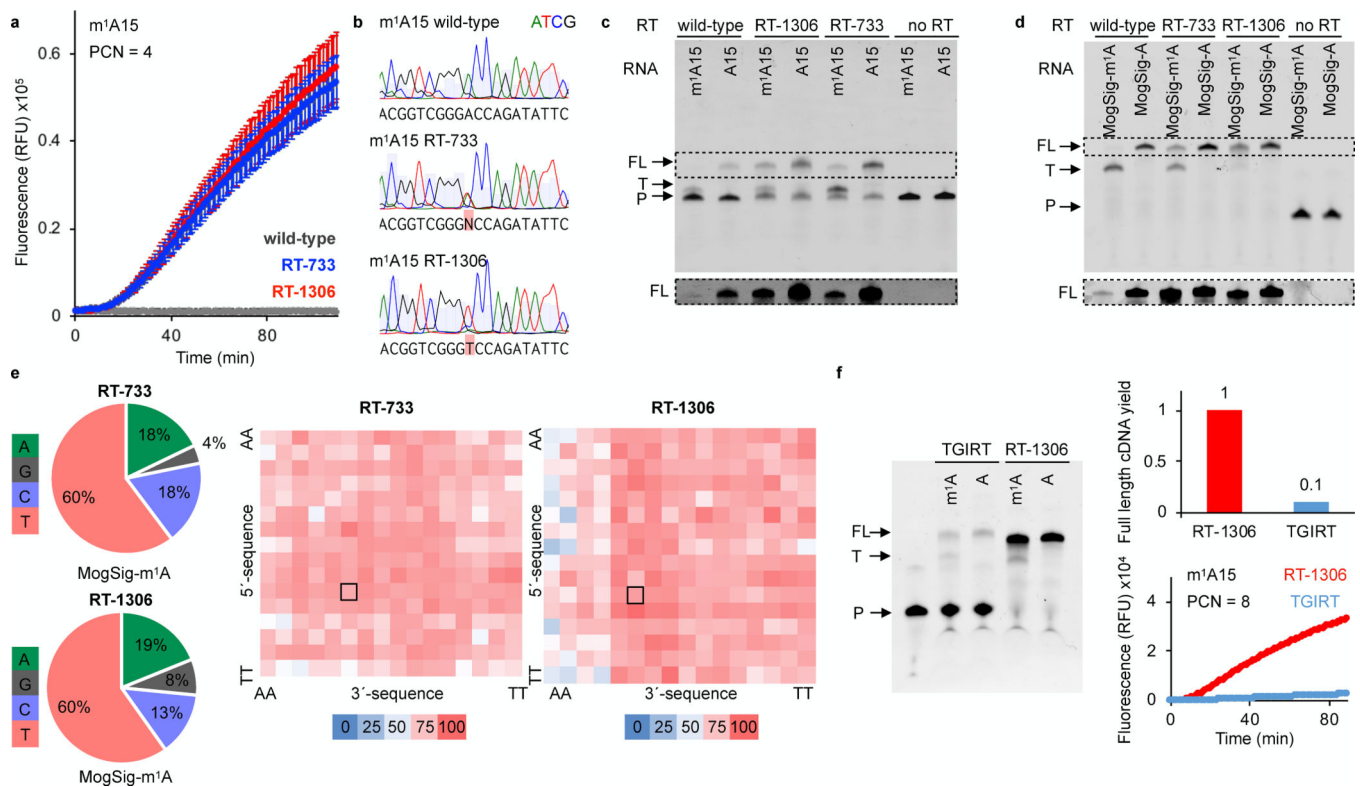
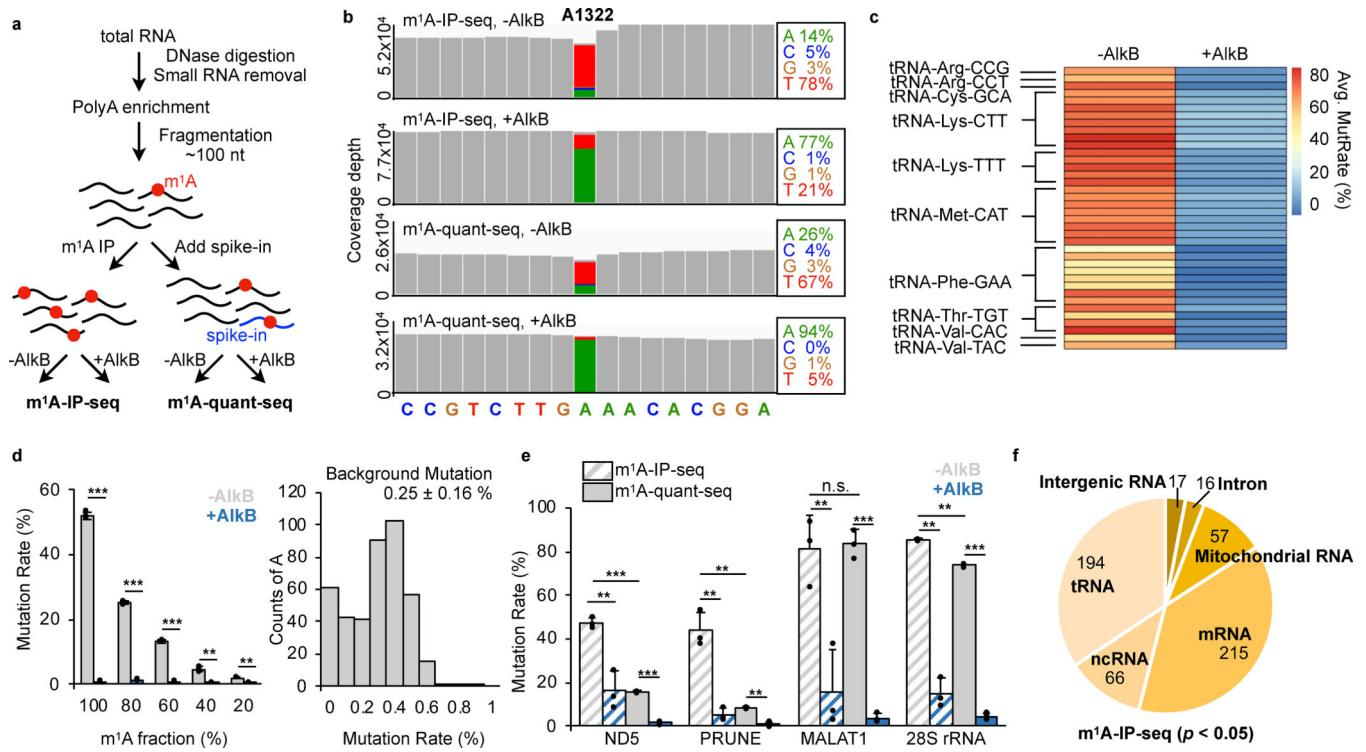


Figure 3.

In vitro characterization of evolved RT variants. **(a)** RT-PCR-IVT assay with purified RT variants with PCN = 4 on the m¹A15 RNA substrate, with mean±s.e.m, from n = 5 independent assays. **(b)** Sanger sequencing of the cDNA products from the wild-type and evolved RT variants of the m¹A15 RNA substrate. Mutations at the m¹A position are highlighted. **(c)** Read-through assay results for the purified RT variants on the m¹A15 RNA and control A15 RNA. RT primer, truncated and full-length cDNA products are labeled with “P”, “T” and “FL”, respectively. The portion of the gel containing the “FL” bands are presented under over-exposure to facilitate visualization of products made by wild-type RT. **(d)** Read-through assay results shown are performed on the synthetic oligonucleotide libraries ModSig-m¹A and ModSig-A. Abbreviations and over-exposed gel picture follow those in panel c. **(e)** Overall mutation patterns of RT-733 and RT-1306 over ModSig-m¹A RNA are shown by the pie charts. The heatmaps present the sequence-context-dependent mutation rates at the m¹A site by NGS. **(f)** Read-through assay for comparing RT activities of RT-1306 and TGIRT using the ModSig-m¹A (“m¹A”) and ModSig-A (“A”) as substrates. Gel image is overexposed to facilitate visualization of RT products of TGIRT; the full-length cDNA yield was quantified based on of band intensities with ImageJ and shown by the bar graph normalized to the full-length cDNA yield by RT-1306. Lower panel shows the RT-PCR-IVT assay data against m¹A15 RNA by TGIRT and RT-1306 with PCN=8.

**Figure 4.**

Single-base resolution m¹A-IP-seq and m¹A-quant-seq. **(a)** Experimental procedure of m¹A-IP-seq and m¹A-quant-seq. **(b)** IGV coverage traces of mutation signatures of m¹A1322 in cytosolic 28S rRNA by m¹A-IP-seq and m¹A-quant-seq. **(c)** Mutation rates detected on m¹A58 over 38 tRNA genes by m¹A-IP-seq are shown via a heatmap of averaged mutation rates from $n = 3$ biological replicates. **(d)** m¹A mutation rates without (in grey) and with (in blue) AlkB treatment are shown for the spike-in sample by m¹A-quant-seq on the left. Error bars represent s.d. from $n = 3$ biological replicates with individual data point shown by the overlaid dots. One-sided t-test is performed for changes in m¹A sites mutation rates upon AlkB treatment; ** $p < 0.01$ and *** $p < 0.001$. Shown on the right is the histogram of mutation rates for all A residues in the spike-in sample, of which mean \pm s.d. is $0.25 \pm 0.16\%$. **(e)** Manual inspection of m¹A sites in the mitochondrial mRNA ND5 (chrM:13711), PRUNE mRNA (chr1: 150980982), lncRNA MALAT1 (chr11: 65273630), and m¹A1322 in 28S rRNA. One-sided t-test is performed for changes in mutation rates upon AlkB treatment, or upon anti-m¹A antibody enrichment; ** $p < 0.01$, *** $p < 0.001$ and n.s. for $p > 0.1$. Error bars represent s.d. of mutation rates from $n = 3$ biological replicates with individual data point shown by the overlaid dots. **(f)** Distribution of RNA types for m¹A sites identified by m¹A-IP-seq with $p < 0.05$ by the beta binomial regression test.

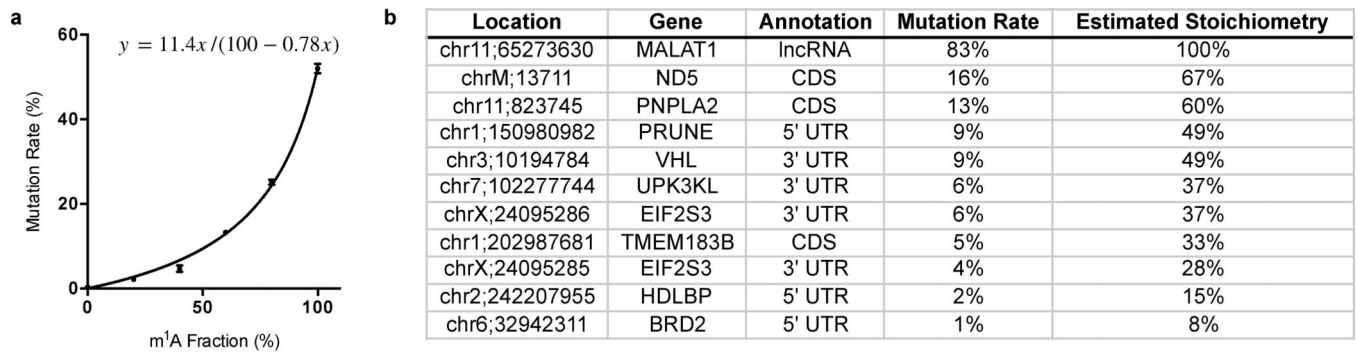


Figure 5.

Estimation of m¹A stoichiometry by m¹A-quant-seq. **(a)** Calibration curve of observed mutation rates versus m¹A fraction based on the spike-in sample. Error bars represent s.d. of mutation rates from n = 3 library replicates. **(b)** Estimation of m¹A stoichiometry for representative m¹A sites in mRNA and lncRNA. Mutation rate is averaged from biological triplicates in the m¹A-quant-seq and estimated stoichiometry is calculated based the calibration curve shown in panel **a**.