Evolution of a reverse transcriptase to map N1-methyladenosine in human messenger RNA

Huiqing Zhou1,2, Simone Rauch1,3, Qing Dai1,4, Xiaolong Cui1,2, Zijie Zhang1,2,3,4, Sigrid Nachtergaele1,2,4, Caroline Sepich1,2,3,5, Chuan He1,2,3,4, and Bryan C. Dickinson1,2,3,5

Chemical modifications to 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 platform for directed evolution 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 N1-methyladenosine (m1A). Iterative rounds of selection yielded reverse transcriptases with both robust read-through and high mutation rates at m1A sites. The optimized evolved reverse transcriptase enabled detection of well-characterized m1A sites and revealed hundreds of m1A sites in human mRNA. This work develops and validates the reverse transcriptase evolution platform, and provides new tools, analysis methods and datasets to study m1A biology.

Enrichment methods (for example, m6A sequencing4, methylated reading signatures (that is, mutations or stops) specifically introduced into RNA samples) are high-throughput approaches that typically rely on next-generation sequencing (NGS) to identify RNA modifications 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 N1-methyladenosine (m1A). Iterative rounds of selection yielded reverse transcriptases with both robust read-through and high mutation rates at m1A sites. The optimized evolved reverse transcriptase enabled detection of well-characterized m1A sites and revealed hundreds of m1A sites in human mRNA. This work develops and validates the reverse transcriptase evolution platform, and provides new tools, analysis methods and datasets to study m1A biology.

Modifications of mRNA play critical regulatory roles in a range of mammalian biological processes, such as cell differentiation, sex determination and development1,2, and are dysregulated in a number of human diseases3. RNA modifications, such as N6-methyladenosine (m6A) and m1A have been mapped on a transcriptome-wide scale using antibody-based enrichment methods (for example, m6A sequencing4, methylated RNA immunoprecipitation sequencing5, m1A sequencing6 and m1A-RNA immunoprecipitation with demethylase-assisted RNA sequencing7,8). Although enrichment-based approaches are powerful tools 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.

Techniques that offer single-base resolution and that allow detection of the exact location and the modification stoichiometry on a transcript are rapidly being developed9-12. Base-resolution methods using next-generation sequencing (NGS) typically rely on reading signatures (that is, mutations or stops) specifically introduced at the modification site by a reverse transcriptase (RT). Notably, mutations are more advantageous over stops because the latter is subject 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 example, RTs were engineered to facilitate read-through of secondary structures of RNA with improved thermostability13,14, repurposed from DNA polymerases to harness their proof-reading activity14, and evolved to recognize modifications, such as N6-methyldeoxyxosine15 and m1A15. However, few engineered RTs were applied successfully in transcriptome-wide mapping of targeted modifications. We reason that this methodological paucity can be attributed to limitations in the 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 NGS14,15.

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 basepairing (Fig. 1a). Although m1A has long been known to be present in transfer RNAs and ribosomal RNAs10,18–20, mass spectrometry and transcriptome-wide m1A profiling studies21 revealed that m1A also occurs in mRNA and long non-coding RNA (lncRNA). Recently, two base-resolution mapping studies of m1A11,21 offered inconsistent findings regarding the number of reported m1A sites in human mRNA22,23. Among many reasons for such discrepancies, the thermostable group II intron reverse transcriptase (TGGRT)11 used to read through m1A exhibits low activity in primer incorporation24. Such low RT activity might render it insensitive and incapable of generating reproducible m1A maps, especially for m1A sites with 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 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 HEK-293T-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 applied the evolved RT in ‘m1A-quant-seq’, which allowed us to estimate m1A stoichiometries at individual sites in the transcriptome.
Results
Choice of reverse transcriptase 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’ adducts. Although the native HIV-1 RT functions as a heterodimer containing two subunits, a 66-kDa catalytic domain (p66) and a 51-kDa structure-stabilizing domain (p51), it was shown to be active as a homodimer with only the catalytic subunit. Additionally, the high expression level of p66 in *Escherichia 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-1T. To detect modifications on A, we first found a U-to-A mutation at U15 in the Broccoli sequence that completely abolishes the fluorescence (Fig. 1a, Supplementary Notes). Therefore, when an RT generates truncations (Fig. 1c,d and Supplementary Fig. 2a), it was shown to be active as a homodimer with only the catalytic subunit. Additionally, the high expression level of p66 in *Escherichia coli* facilitates screening efforts. Therefore, we chose to evolve the p66 subunit of HIV-1 RT.

To detect such RT mutations, we designed an ‘RT-PCR-IVT’ assay in which reverse transcription of m’A 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 can generate a truncated or full-length cDNA product in presence of m’A (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).

Directed evolution of HIV-1 reverse transcriptase. We validated that the RT-PCR-IVT assay functions with using enzymes expressed in crude cell lysates, which further facilitated rapid screening (Supplementary Notes and Supplementary Table 1) as positive and negative RNA controls for 100% and 0% reactions using U15 and A15 RNA (sequences in Supplementary Table 1). It was shown to be active as a homodimer with only the catalytic subunit. Additionally, the high expression level of p66 in *Escherichia coli* facilitates screening efforts. Therefore, we chose to evolve the p66 subunit of HIV-1 RT.

Performing the assay with the m’A15 substrate and wild-type HIV RT resulted in negligible fluorescence, predominantly because of RT truncations (Fig. 1c,d and Supplementary Fig. 2a).

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). We performed the RT-PCR-IVT assay without the intermediate purification steps and validated the products of each step using gel electrophoresis (Fig. 1d and Supplementary Notes). Performing the assay with the m’A15 substrate and wild-type HIV RT resulted in negligible fluorescence, predominantly because of RT truncations (Fig. 1c,d).
After growth of cell cultures with induced over-expression of RT variants in the 96-deepwell plate, cell lysates were collected from the template selectivity. In our case, the six favorable amino acid mutation sites located on both sides of the replication base pair (PDB ID: 1RTD) and might contribute to decreased geometric constraints at the active site, 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 mutations: D76A, R78K, W229Y, M230L, V75F, F77A (Fig. 2c and Supplementary Fig. 2d). We saturated the key sites to find optimal mutations, focusing on amino-acid positions 76 and 78 in the second round (Libraries 9–11), and positions 229 and 230 in the third round (Libraries 12 and 13). Results suggested that the most promising variant contained the mutations D76A, R78K, W229Y, M230L, V75F and F77A (Supplementary Notes). In the third round (Libraries 14–17), we also screened sites L74, V75, F77, E79 and L80, which surround positions 76 and 78 (Supplementary Table 2). RT-973 (with mutations V75F and F77A) yielded further improvements in fluorescence signal (Fig. 2c).

Previous studies of polymerase engineering and structural characterizations suggest mutations altering interactions between domains around the active site of the polymerase might affect the template selectivity. In our case, the six favorable amino acid mutation sites located on both sides of the replication base pair (PDB ID: 1RTD) and might contribute to decreased geometric constraints at the active site, 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 mutations: D76A, R78K, W229Y, M230L, V75F, F77A (Fig. 2c and Supplementary Table 1). We purified RT-1306, and additionally RT-733 as one representative variant along the selection, for further biochemical characterization.

Biochemical characterization of evolved reverse transcriptases. The fluorescence response of the RT-PCR-IVT assay using both purified RT-733 and RT-1306 enzymes (Supplementary Fig. 3a) was more than 40-fold higher than that of 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 m1 A to T mutation signature at the m1 A at PCN = 15 for RT-1306 containing mutations: D76A, R78K, W229Y, M230L, V75F and F77A (Fig. 2c and Supplementary Fig. 2d). We saturated the key sites to find optimal mutations, focusing on amino-acid positions 76 and 78 in the second round (Libraries 9–11), and positions 229 and 230 in the third round (Libraries 12 and 13). Results suggested that the most promising variant contained the mutations D76A, R78K, W229Y, M230L, V75F and F77A (Supplementary Notes). In the third round (Libraries 14–17), we also screened sites L74, V75, F77, E79 and L80, which surround positions 76 and 78 (Supplementary Table 2). RT-973 (with mutations V75F and F77A) yielded further improvements in fluorescence signal (Fig. 2c).

Previous studies of polymerase engineering and structural characterizations suggest mutations altering interactions between domains around the active site of the polymerase might affect the template selectivity. In our case, the six favorable amino acid mutation sites located on both sides of the replication base pair (PDB ID: 1RTD) and might contribute to decreased geometric constraints at the active site, 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 mutations: D76A, R78K, W229Y, M230L, V75F, F77A (Fig. 2c and Supplementary Table 1). We purified RT-1306, and additionally RT-733 as one representative variant along the selection, for further biochemical characterization.

Biochemical characterization of evolved reverse transcriptases. The fluorescence response of the RT-PCR-IVT assay using both purified RT-733 and RT-1306 enzymes (Supplementary Fig. 3a) was more than 40-fold higher than that of 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 there was an A-to-T mutation signature at the m1A site for RT-1306 and RT-733, rather than the wild-type RT (Fig. 3b,
In vitro characterization of evolved RT variants. a, RT-PCR-IVT assay with purified RT variants with PCN = 4 on the m1A15 RNA substrate, as 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 site position are highlighted. c, Read-through assay results for the purified RT variants on the m1A5 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 overexposure to facilitate visualization of products made by wild-type RT. d, Results from read-through assay performed on the synthetic oligonucleotide libraries ModSig-m1A and ModSig-A. Abbreviations and overexposed gel pictures are as described in c, e. Overall mutation patterns of RT-733 and RT-1306 over ModSig-m1A RNA are shown as pie charts. The heatmaps present the sequence-context-dependent mutation rates at the m1A site by NGS. Mutation rates in percentage are color coded. 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 on the basis of band intensities with ImageJ and is shown by the bar graph normalized to the full-length cDNA yield by RT-1306. The lower panel shows the RT-PCR-IVT assay data against m1A15 RNA by TGIRT and RT-1306 with PCN = 8.

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 than 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 RNA used in the screen, we used an oligonucleotide library of 45-mer synthetic RNAs that differed in sequence from m1A15 and contained a single m1A (‘ModSig-m1A’) flanked by 2-nucleotide random sequences (‘-NNm1ANN-’), where N = A, G, C or U (Supplementary Table 1). The control library ‘ModSig-A’ contains the same sequences but with unmodified A residues (‘-NNANN-’). The read-through assay data from these libraries (Fig. 3d) showed that the wild-type RT predominantly produces a truncated cDNA product, whereas RT-1306 consistently showed 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 site in varying sequence contexts, we performed high-throughput sequencing of cDNA libraries generated from the synthetic ModSig RNA libraries using the wild-type RT, RT-733 and RT-1306. The sequencing results revealed a moderate increase in the 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 dramatically enhanced A-to-T mutation rate (36% for wild-type versus 60% for RT-733 and RT-1306), which is the mutation signature that generated fluorescence signal during the directed-evolution screening. Most notably, this elevated mutation rate was reasonably consistent across the 256 different sequence contexts, and there was 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 improved read-through efficiency, the mutational signatures of the evolved RT variants are largely increased, without substantially affecting the background mutation rate (Supplementary Fig. 4). RT-733 and RT-1306 present comparable overall mutation rates as determined using the sequencing data, whereas RT-1306 possesses a higher read-through efficiency. In vitro comparisons of RT-1306 and TGIRT revealed that RT-1306 yielded 10-fold higher level of full-length cDNA product and a substantially higher ratio of read-through to truncation product than that of TGIRT (Fig. 3f and Supplementary Notes). Together, these data suggest that RT-1306 might be capable of detecting m1A in biological RNA samples.
Mutation signatures in m'A-IP-seq and m'A-quant-seq. We applied RT-1306 in NGS libraries for m'A mapping in mRNA. For m'A-IP-seq, which involves immunoprecipitation (IP) with an anti-m'A antibody, we used a protocol similar to the previously described ‘m'A-MAP’ protocol with optimized library-construction steps (Supplementary Fig. 5 and Supplementary Notes). For m'A-quant-seq, we did not perform m'A-IP; rather, we spiked in synthetic m'A oligonucleotides with various m'A fractions (‘spike-in RNA’) to use for estimating m'A stoichiometry (Fig. 4a and Methods). m'A-IP-seq and m'A-quant-seq protocols yielded high-quality libraries with the expected library length (~250 bp), high alignment rates 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 m'A mutation signatures, we first manually examined the well-characterized m'A sites in rRNA and tRNA. The mutation rates at m'A1322 in 28S rRNA are 78% and 67%, which decrease to 21% and 5% after AlkB treatment, as measured by m'A-IP-seq and m'A-quant-seq, respectively (Fig. 4b). Interestingly, the mutation signatures at m'A showed predominantly an A-to-T mutation (Fig. 4b), different from the predominant A-to-G mutation that was observed at the same site by the TGIRT. We observed robust mutation signatures (75 ± 11%; mean ± s.d.) for m'A58 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 were also consistently observed in the spike-in m'A 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 ± 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 ± 0.01%), tRNA (0.1 ± 0.04%) and GAPDH mRNA (0.6 ± 0.01%). Taken together, these reference sites showed that the evolved RT-1306 can generate high and accurate mutation rates at m'A sites in biological RNA samples with low levels of background mutations at A sites.

Mutations and read-through of m'A sites in mRNA and IncRNA. We examined mutation signatures in our sequencing datasets for three commonly found m'A sites in prior single-base m'A mapping data: PRUNE (annotated as both mRNA and ncRNA in the RefSeq database); MALAT1 (IncRNA); and ND5 (mitochondrial mRNA). For these examined sites, AlkB treatment shows a >50% decrease in mutation rate in both m'A-IP-seq and m'A-quant-seq (P < 0.01, one-sided t-test shown in Fig. 4e). m'A-IP significantly enriches mutation rates for sites in PRUNE and ND5, but not in MALAT1 (Fig. 4e). This is not unexpected, because m'A-IP can enrich the m'A fraction for sites with low m'A stoichiometry,
leading to increased mutation rates in IP libraries as compared to non-IP libraries. Sites with high mA stoichiometry, such as the MALAT1 site, may generate maximal mutation rates that will not further increase by 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 mA sites from mA-IP-seq. Because mA was found as a relatively rare modification in mRNA and IncRNA (0.014 ± 0.004% mA/A in polyA-enriched RNA from HEK293T cells, measured by liquid chromatography–tandem mass spectrometry (LC–MS/MS); Supplementary Fig. 5a), we used mA-IP-seq to identify mA sites with increased sensitivity. We validated the anti-mA antibody activity in vitro by dot blotting and LC–MS/MS quantification following immunoprecipitation (Supplementary Fig. 8 and Supplementary Notes). With mA-IP-seq data, we confirmed the enrichment activity of the antibody by manual inspection and Lorenz curves analysis (Supplementary Fig. 8). Although these tests suggest that mA antibody can enrich mA-containing RNA fragments at examined sites, it remains challenging to evaluate its specificity for all mA sites in the complex context of the human transcriptome. To identify mA sites, we relied on the mutation rate (≥2%) and its sensitivity to AlkB treatment evaluated by beta-binomial regression test over three biological replicates (Supplementary Fig. 9 and Methods). The mutation signatures will help rule out most interfering signals from non-specific antibody-binding behavior.

In total, we identified 565 statistically significant mA sites (P < 0.05; Fig. 4f) from mA-IP-seq; these include 215 mRNA and 66 ncRNA sites (50 annotated as IncRNA in the NONCODE v5 database19), 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 captured by RT-1306 in human transcriptome. To identify mA sites, we relied on the mutation rate (<3%) suggesting that they likely have low mA stoichiometry (Supplementary Fig. 10b).

Estimation of mA stoichiometry using mA-quant-seq. We applied the same beta binomail statistical test in processing mA-quant-seq data. mA-quant-seq identified 55 potential mRNA and IncRNA mA 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 mA fraction to fit the values measured from the spike-in sample (Fig. 5a). We show an estimation of the mA fraction for a subset of ten sites with a relatively high mutation rate and a small P value (that is, high sensitivity to AlkB treatment), identified by mA-quant-seq based on the calibration curve (Fig. 5b). The list of genes that likely possess mA at high and low fractions without antibody enrichment can serve as a useful reference list for future studies of mA biology. Encouragingly, mutation signatures and AlkB sensitivity observed by mA-quant-seq are highly comparable to those captured by a targeted library approach with deep coverage for PRUNE and MALAT1 mA sites (Supplementary Fig. 11b,c), suggesting that RT-1306 is applicable in locus-specific mA detection. We believe this method can be used to quantify and compare mA levels in different biological contexts, such as stress or disease.

Discussion

Our data suggest that mA could be present in a large number of sites in human mRNA, although it might occur with low stoichiometry, especially for those we could only detect with mA-IP-seq. There could be new mA-specific mRNA methyltransferases yet to be discovered, or alternatively, mA sites could be installed by moonlight activities of tRNA modification enzymes. Previous studies have already revealed preferential mA modification in mitochondrial mRNA11,21. Context-dependent RNA modifications is an emerging theme in transcriptome regulation32.

Biological RNAs vary widely in endogenous abundance and 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 mA. 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 by comparing RT-1306 with TGIRT, data reproducibility becomes worse when considering both immunoprecipitation and demethylation treatment steps24,25 (Supplementary Fig. 12 and Supplementary Notes). Our antibody-free mA-quant-seq reduced

**Table 4.** Mutation signatures of sites are captured by RT-1306 in human transcriptome. To identify mA sites, we relied on the mutation rate (<3%) suggesting that they likely have low mA stoichiometry (Supplementary Fig. 10b).

| Location | Gene | Annotation | Mutation rate | Estimated stoichiometry |
|----------|------|------------|---------------|-------------------------|
| chr1:65273630 | MALAT1 | IncRNA | 8% | 100% |
| chr1:13711 | ND5 | CDS | 16% | 67% |
| chr11:823745 | PNPLA2 | CDS | 13% | 60% |
| chr:150980982 | PRUNE | 5 UTR | 9% | 49% |
| chr:2:10194784 | VHL | 3 UTR | 9% | 49% |
| chr:7:102277744 | UPIK3BL | 3’ UTR | 6% | 37% |
| chr:2:24095286 | EIF2S3 | 3’ UTR | 6% | 37% |
| chr:1:202987681 | TMEM183B | CDS | 5% | 33% |
| chr:2:24095285 | EIF2S3 | 3’ UTR | 4% | 28% |
| chr:2:242207955 | HDLBP | 5 UTR | 2% | 15% |
| chr:6:32942311 | BRD2 | 5 UTR | 1% | 8% |

**Fig. 5 | Estimation of mA stoichiometry by mA-quant-seq.** a, Calibration curve of observed mutation rates versus mA fraction based on the spike-in sample. Error bars represent the s.d. of mutation rates from n = 3 library replicates. b, Estimation of mA stoichiometry for representative mA sites in mRNA and IncRNA. The mutation rate is averaged from biological triplicates in the mA-quant-seq, and estimated stoichiometry was calculated based the calibration curve shown in a.
such variance and enabled us to more directly estimate single-site m’A stoichiometry.

Although mutation-based sequencing approaches have advanced the quality of mapping methods and identification of potentially modified sites, there are concerns associated with peak calling as summarized by Sas-Chen et al.10. 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 the relative confidence level for each reported site. We do not rule out that detected sites can be mis-identified owing to error-prone sequence contexts such as homopolymeric sequences, or ambiguous assignments coming from repeated sequences in the transcriptome, such as MT-RNR2-like genes. For such incidents, cross validation by orthogonal experiments or long-read sequencing will be required to confirm the presence and fraction of m’A.

Noting that our m’A modification fraction detection limit is roughly 8% (corresponding to a 1% mutation rate), we likely underestimated the number of m’A sites with a low modification stoichiometry. The current calibration function 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 that are 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, 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′-cap.

The evolved RTs will also find general applications beyond m’A-seq of mRNA, for example, in mRNA sequencing and in DMS-seq for probing of RNA secondary structure11,12, where m’A is the reaction product of more single-stranded A residues in RNA reacted with dimethyl sulfoxide (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 for the screen. Additionally, for modifications that do not perturb base pairing as much as m’A, we anticipate that larger library sizes will be needed to uncover sensitive RT enzymes. However, the robustness of the screening approach should make it amenable to automation, allowing us to screen RT libraries of much sizes, which we are currently pursuing.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0550-4.

Received: 3 December 2018; Accepted: 5 August 2019; Published online: 23 September 2019

References
1. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA modifications in gene expression regulation. Cell 169, 1187–1200 (2017).
2. Eyre, M., Harada, B. T., Behm, M. & He, C. RNA modifications modulate gene expression during development. Science 361, 1346 (2018).
3. Jonkhout, N. et al. The RNA modification landscape in human disease. RNA 23, 1754–1769 (2017).
4. Dominissini, D. et al. Topology of the human and mouse m6A RNA methylosomes revealed by m6A seq. Nature 485, 201 (2012).
5. Meyer, K. D. et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. Cell 149, 1635–1646 (2012).
6. Dominissini, D. et al. The dynamic N1-methyladenosine methylome in erythrocytic messenger RNA. Nature 530, 441 (2016).
7. Li, X., Xiong, X. & Yi, C. Epitranscriptome sequencing technologies: decoding RNA modifications. Nat. Methods 14, 23 (2016).
8. Li, X. et al. Transcriptome-wide mapping reveals reversible and dynamic N1-methyladenosine methylome. Nat. Chem. Biol. 12, 311 (2016).
9. Zheng, G. et al. Efficient and quantitative high-throughput mRNA sequencing. Nat. Methods 12, 835 (2015).
10. Cozen, A. E. et al. ARM-seq: AlkB-facilitated RNA methylation reveals a complex landscape of modified mRNA fragments. Nat. Methods 12, 879 (2015).
11. Li, X. et al. Base-resolution mapping reveals distinct m1A methylome in nuclear- and mitochondrial-encoded transcripts. Mol. Cell 68, 993–1005.e1009 (2017).
12. Schwartz, S. & Motorin, Y. Next-generation sequencing technologies for detection of modified nucleotides in RNAs. RNA Biol. 14, 1124–1137 (2017).
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).
14. Zhao, C., Liu, F. & Pyle, A. M. An ultraprospective, accurate reverse transcriptase encoded by a metazoon group II intron. RNA 24, 183–195 (2018).
15. Ellefson, J. W. et al. Synthetic evolutionary origin of a proofreading reverse transcriptase. Science 352, 1590 (2016).
16. Huber, C., von Watzdorf, J. & Marx, A. 5-methylcytosine-sensitive variants of Thermococcus kodakarenensis DNA polymerase. Nucleic Acids Res. 44, 9881–9890 (2016).
17. Harcourt, E. M., Ehrenschwender, T., Batista, P. J., Chang, H. Y. & Kool, E. T. Identification of a selective polymerase enables detection of N6-methyladenosine in RNA. J. Am. Chem. Soc. 135, 19079–19082 (2013).
18. Helm, M., Giege, R. & Florentz, C. A. Watson–Crick base-pair-disrupting methyl group (m1A9) is sufficient for cloverleaf folding of mitochondrial RNA(lys). Biochemistry 38, 13338–13446 (1999).
19. Anderson, J. T. & Droogmans, L. in Fine-Tuning of RNA Functions by Modification and Editing (ed Grossenbacher, H.) 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).
21. Safra, M. et al. The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. Nature 551, 251 (2017).
22. Xu, L. et al. Three distinct 3-methylcytidine (m3C) methyltransferases modify tRNA and mRNA in mice and humans. J. Biol. Chem. 292, 14695–14703 (2017).
23. Patterson, J. T., Nickens, D. G. & Burke, D. H. HIV-1 reverse transcriptase pausing at bulky 2′-adects is relieved by deletion of the nase H domain. RNA Biol. 3, 163–169 (2006).
24. Kholostova, L. A., Wang, J., Friedman, J. M., Rice, P. A. & Steitz, T. A. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science 256, 1783 (1992).
25. Huang, H., Chopra, R., Verch, E. & Harrison, S. C. Structure of a Covalently Trapped Catalytic Complex of HIV-1 Reverse Transcriptase: Implications for Drug Resistance. Science 282, 1669 (1998).
26. Müller, B. et al. Co-expression of the subunits of the heterodimer of HIV-1 reverse transcriptase in Escherichia coli. J. Biol. Chem. 264, 13975–13978 (1989).
27. Filonov, G. S., Moon, J. D., Swensen, N. & Jeffrey, S. R. Broccoli: rapid selection of an RNA mimic of green fluorescent protein by fluorescence-based selection and directed evolution. J. Am. Chem. Soc. 135, 18637–18643 (2013).
28. Chen, T. et al. Evolution of thermophilic DNA polymerases for the recognition and amplification of C2-modified DNA. Nat. Chem. 8, 556 (2016).
29. Betz, K. et al. Structural insights into dna replication without hydrogen bonds. J. Am. Chem. Soc. 135, 18637–18643 (2013).
30. Diaz, A., Nellore, A. & Song, J. S. CHANCE: comprehensive software for quality control and validation of ChiP-seq data. Genome Biol. 13, R98 (2012).
31. Fang, S. et al. NONCODEv5: a comprehensive annotation database for long non-coding RNA. Nucleic Acids Res. 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. Mol. Cell 74, 640–650 (2019).
33. Dominissini, D. & Rechavi, G. Loud and clear epitranscriptomic m1A signals: now in single-base resolution. Mol. Cell 68, 823–826 (2017).
34. Xiong, X., Li, X., Wang, K. & Yi, C. Perspectives on topology of the human m1A methylome at single nucleotide resolution. RNA 24, 1437–1442 (2018).
35. Sas-Chen, A. & Schwartz, S. Misincorporation signatures for detecting modifications in RNA: Not as simple as it sounds. Methods 156, 53–59 (2019).
36. Zubradt, M. et al. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. Nat. Methods 14, 75 (2016).
37. Guo, J. U. & Bartel, D. P. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. Science 353, (2016).

Acknowledgements
We thank the entire staff at the University of Chicago Comprehensive Cancer Center sequencing facility at for performing the Sanger and NGS sequencing measurements; T. Pan for sharing the ModSig libraries; L. Zhang and C. Liu for sharing the protocols for constructing biological RNA libraries; L. Hu and L. 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 the 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.

Author contributions
B.C.D., C.H., and H.Z. conceived the idea of the project and designed the experiments. H.Z. performed the directed-evolution experiments with assistance from S.R., who performed 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 the 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 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.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0550-4.
Correspondence and requests for materials should be addressed to C.H. or B.C.D.

Peer review information
Rita Strack was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.
Reprints and permissions information is available at www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019
Methods

**Sample preparations.** DNA and RNA oligonucleotides. DNA primers, primers with fluorescein amide (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 m^A^-IP-seq and m^A^-quant-seq were ordered from IDT with high-performance liquid chromatography (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 m^A^ modifications and Dimodification reagent. After deprotection, the RNA oligonucleotides were purified through HPLC with a C18 column and were 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% 8 M urea polyacrylamide gel electrophoresis (PAGE) gel. The 33-mer A15 m^A^-a, and m^A^-bA' showed cleavage at 37 °C and 25 °C, respectively. In samples, 43-mer RNAs showed impurity bands, so we performed gel purification for all these RNAs with 10% PAGE gel and RNA recovery with the ZR small-sample 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 6His-tag on the N-terminus, connected by a GGGS linker. The HIV-1 RT and its evolved variants were overexpressed in E. coli BL21 (DE3) cells and purified following the protocol reported previously\(^38\). Briefly, two liters of cell culture were grown at 37 °C for 3 h in LB medium with 80 μM kanamycin until the optical density at 600 nm (OD\(_{600}\)) reached 0.5–0.6. The overexpression was induced by 0.5–1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and cells were harvested at 37 °C for 2.5 h and then at 16 °C overnight. Cells were harvested and resuspended in 40 mM lysis buffer (50 mM Na\(\text{HPO}_4\), and Na\(\text{H}_2\text{PO}_4\), 0.5 M NaCl, pH 7.8, dissolved half-tablet of the proteasome inhibitor cocktail, Pierce) per liter of culture. The cells were then lysed by sonication and centrifuged at 12,000 rpm for 40 min at 4 °C. Solubilized proteins in the supernatant were first purified using Hi50 Ni Superflow resins (Clontech Laboratories, Inc.) and were eluted with 50 mM to 0.5 M gradient imidazole buffer containing 50 mM Na\(\text{HPO}_4\), and Na\(\text{H}_2\text{PO}_4\), 0.5 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. Buffer A (50 mM Bis-tris pH 7.0) with an 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 95.3% Buffer A and 2.5% Buffer B (50 mM Bis-tris pH 7.0, 1 M 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 eluted with the storage buffer (50 mM Tris-HCl, 25 mM Na\(\text{Cl}\), 1 mM EDTA, 50% glycerol, pH 7.0). Purified proteins were concentrated to 200–300 μL with 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 −1 mg of purified proteins for the wild-type HIV-1 p71.

**Biochemical assays.** Read-through assay. The RT read-through assays were performed in 10-μL reaction volumes containing 1× RT primer and RNA substrate, 4 mM deoxyribonucleoside triphosphate (dTTP) mixture (1 mM each dTTP) and 3 μl crude lymphocyte 1× 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 1× PCR buffer, and 4 μM each forward and reverse PCR primer, with a given number of PCR cycles. PCR reactions were carried out with a 30-s annealing step at 58 °C and 45-s elongation step at 72 °C. Finally, 7 μl of each PCR reaction mix was added as template to a 20-μl in vitro transcription reaction on a 384-well plate with a glass bottom (Cellvis), together with 8 mM ribonucleoside triphosphate (rNTP) mixture (2 mM each rNTP), 50 mM DFBHI-1T (LuceRNA, Inc.) and 0.5 μl purified RT RNA polymerase. The in vitro transcription reactions were monitored by a plate reader (BioTek, Inc.) for 1.5–3 h, at an interval of 1 read per minute, with the excitation and emission wavelengths at 472 nm and 507 nm, respectively.

**LC–MS/MS.** AliKb treatment. AliKb treatment of RNA in vitro was performed in 40-μL reactions containing 10–50 pmol RNA substrate depending on the detection assay, purified AliKb (at least 4 fold of the RNA molarity), and 0.5 μM Sophus RNA Inhibitor in 1× reaction buffer. For initial tests of purified AliKb activity, 1× reaction buffer contained 300 mM KCl, 2 mM MgCl\(_2\), 50 μM BSA, 50 μM (NH\(_4\))\(_2\)Fe(SO\(_4\))\(_3\), 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 AliKb treatment reactions were carried out at 25 °C for 2 h, 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.

**Dot blotting.** 5-mer synthetic RNA oligonucleotides were made into gradient concentrations in 10, 8, 4, 2, 1, and 0.5 ng μl\(^{-1}\), and 1 μl of each was dotted on buffer (ThermoFisher), supplemented with 1 mM MgCl\(_2\), 1X in vitro transcription buffer contained 40 mM Tris-HCl (pH 7.9), 30 mM MgCl\(_2\), 2 mM spermidine, 20 mM DTT.

**Cloning and preparation of crude lysate for RT variants.** Libraries were constructed with 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 to 96-deep-well plates and grown overnight, with shaking, at 37 °C. Overnight cultures were diluted by 15-fold using LB medium with kanamycin antibiotic, and then were grown for 2.5 h at 37 °C. Cells from the rest of the overnight culture cell were 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 h. The cells were then collected by spinning down the deep-well plates at 4,000 rpm for 20 min at 15 °C. The supernatant was discarded by decanting the plate. The cells were kept at −80 °C and were stored immediately before the screening assays. Collected cells were lysed on the deep-well plate by adding 80 μl lysing buffer, which contained 50 mM sodium phosphate, 300 mM NaCl, proteinase inhibitor cocktail (Thermo Scientific) and 1 mg ml\(^{-1}\) lysozyme (Fishier 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 then equilibrated incubated in the shaker at 37 °C for 1 h, and then spun down at 4,000 rpm for 80 min. At 4 °C. We then pipetted out 40 μl cell lysate from the supernatant for the screening assays.

**Screening assays.** The reverse transcription assays with the crude lysate screen were performed in 10-μl volumes containing 0.5μM RT primer and RNA substrate, 4 mM deoxyribonucleoside triphosphate (dTTP) mixture (1 mM each dTTP) and 3 μl crude lysate in 1× 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 1× PCR buffer, and 4 μM each forward and reverse PCR primer, with a given number of PCR cycles. PCR reactions were carried out with a 30-s annealing step at 58 °C and 45-s elongation step at 72 °C. Finally, 7 μl of each PCR reaction mix was added as template to a 20-μl in vitro transcription reaction on a 384-well plate with a glass bottom (Cellvis), together with 8 mM ribonucleoside triphosphate (rNTP) mixture (2 mM each rNTP), 50 mM DFBHI-1T (LuceRNA, Inc.) and 0.5 μl purified T7 RNA polymerase. The in vitro transcription reactions were monitored by a plate reader (BioTek, Inc.) for 1.5–3 h, at an interval of 1 read per minute, with the excitation and emission wavelengths at 472 nm and 507 nm, respectively.
a Hyb-N membrane (GE Healthcare) optimized for nucleic acid transfer. Samples were left on the benchtop for around 5 min to dry completely and then ultraviolet-crosslinked to the membrane by running auto-crosslinking mode twice. The membrane was then blocked with 1x TBST at room temperature; washed with 1x PBST for 4 times with 10 min each time. The membrane was incubated with α-m antibody 0.1 g ml⁻¹ 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 h at room temperature. The membrane was washed with 1x PBST 3 times and water-developed using 1:5 SuperSignal West Pico PLUS substrate and peroxide solution, then imaged by the FluorChem R imager under chemiluminescence detection.

m⁵A immunoprecipitation for LC–MS/MS. Total RNA extracted from HEK293T cells with Trizol reagent was first treated with DNase I, and then small RNA fraction was purified using the MEGAcear kit. Remaining total RNA was then fragmented into fragments of 100–150 nucleotides in length by heating at 94°C for 5 min 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 Oligo disappear kit (Zymo Research). 40 μg of fragmented RNA was used as 'Input' RNA for m⁵A IP. RNA was incubated with 10 μg of m⁵A 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 h. 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 h while shaking. Each IP was washed with 1 x PBST buffer, followed by 1 x PBST buffer (10 mM Tris-HCl pH 7.4, 1 mM EDTA, and 0.25% SDS) at 37°C for 1 h while shaking. The elution mixture was further purified by the Oligo disappear 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-mA 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 manufacturer’s instructions for all steps with the following changes. First, we used customized 5′ and 3′ ligation adaptors containing an additional NNNNNN randomized sequence as well as a unique barcoding sequence (Supplementary Table 1) in order to alleviate sequence biases during ligation and to 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 an evolved RT variant. The RT reaction was performed following both 5′ and 3′ ligation adaptors, directly followed by PCR amplification; therefore, read-through cDNA products rather than truncated cDNAs were amplified and sequenced.

Biological RNA libraries. Step-by-step protocols for m⁵A-IP-seq and m⁵A-quant-seq are described in the Extended Protocols in the Supplementary Notes. Purified cDNA was PCR amplified with a ‘+AlkB’ IP RNA. The IP RNA samples without (‘–AlkB’) and with (‘+AlkB’) AlkB treatment were subjected to library constructions procedures separately, following the published m⁵A-Map ligation-based protocol with optimized purification procedures for the RNA samples (Fig. S3B and Extended Supplementary Notes). For the m⁵A-quant-seq, we added 1 ng of spike-in RNA into 10 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 the Extended Protocols for m⁵A-IP-seq (Supplementary Notes), and was made into +AlkB libraries, the rest of RNA fragments with spike-in were used for the –AlkB library. Final library sizes for all replicates around 220–230 bp are shown in the bioanalyzer data (Supplementary Fig. 6).

Targeted library. RNA was purified as for m⁵A-IP-seq, until the fragmentation step. –300 ng of polyA RNA fragments were then subjected to AlkB treatment and purification procedures described. The RNA was treated with PKN for end repair, ligated with a 3′-adaptor, reverse transcribed into cDNA with RT-1066 and purified as in the m⁵A-IP-seq library approach (Extended Protocols in the Supplementary Notes). Purified cDNA was PCR amplified with locus-specific PCR primers (Supplementary Table 1). For each targeted site, individual PCR was performed by using RNA as a template. PCRs 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’ program with the option ‘-q 20 -m 30’ to filter reads by the sequencing quality >20 and a minimum read length of 30. As both ligation adaptors contain a random 5-mt 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 by the fastx_collapser program. The resulting cleaned data were further analyzed by the in-house python scripts ‘countmut.py’ to count mutations at the m⁵A 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 the unique molecular identifier (UMI) in the R1 reads as previously noted in m⁵A-Map processing. The flow chart summarizing the pipeline is summarized in Supplementary Fig. 9. Raw data (R2 reads of the pair-end data) were first processed with the ‘clumpify.sh’ program in the BBMap package with the option ‘dedupe sub=0’ to remove PCR duplicates. Adapters were trimmed by the ‘cutadapt’ program, while reads were filtered by quality and length with options ‘-a AGACTGGAGAAGCTCTGTTAGAGGAAAGATTAGATCTCCGGTGTCGCCGTATCATT -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 has PSIaligner protocol with the option ‘--trim5 11’ and UMI sequence was removed with the script ‘run_cut_alignment_R2_hg19.sh’, and the default ‘soft-clipping’ option was applied with the alignment. We found that turning on this option was crucial to rule out non-templated addition of RT at the ends, which can lead to high-probability false-positive mutation rates, although we do not rule out that having this option on would likely decrease the sensitivity in detecting 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 with the 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 filtered 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 8,000 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 a mutation rate of 10–100× by at least 5 reads for more than 50 sites. The bam-readcount output files of the –AlkB libraries with the script ‘parse_pos.py’ or the script ‘parse_neg.py’; corresponding sites were then found accordingly in the +AlkB library via the script ‘control_Parse_pos.py’ or ‘control_Parse_neg.py’. Finally, we used ‘clean_unread.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 m⁵A identification using B scripts.

Base-resolution m⁵A identification in m⁵A-IP-seq and m⁵A-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 ‘CallowA.ipr’ or ‘CallowA.input.R’ script for m⁵A site identification based on defined criteria, gene annotation and statistical tests. m⁵A 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 replicates; (2) sites that possess at least 1 mutation
rates in –AlkB libraries; (3) sites that are sensitive to AlkB treatment in three biological replicates, that is, Mutation Rate (−AlkB) − Mutation Rate (+AlkB) > 0. 2,041 sites were selected under these criteria for mA-IP-seq, and then annotated by the ChiPseeker38 and GenomicFeatures39 packages in R. Single RNA sites were annotated following the priority list: (1) the single nucleotide polymorphism collected for HEK293T cells62; (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/software/igemone.html); (5) long non-coding RNA from the NONCODE v5 database90; (6) other Ref-seq-annotated non-coding RNA according to the ‘hg19_UCSC.gtf’; (7) integrigenic RNA for the rest of the non-annotated sites. We also crosschecked the resulting sites with the RNA editing database rigorously annotated database of A-to-I RNA editing (RADAR)39 for sites that are potentially assigned as RNA editing rather than mA. In principal, 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 mA sites that mis-assigned as editing. The resulting 2,041 sites were tested by the beta-binomial regression49 and reported in Supplementary Table 4 with P values. mA-quant-seq data were processed following the same pipeline, except using a cut-off mutation count of 3 mutation 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-IRNAs.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 rational with the ChIP-Seq quality control tool CHANCE40. 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 Bedtools41. To eliminate the impact of different genes expression, we have normalized the reads counts in IP samples with input samples. The top 100,000 transcriptomic bins were used to generate Lorenz Curve.

Enrichment peak calling by McRIP tools. We performed IP enrichment peak calling based on enrichment of read count in mA-IP-seq over mA-quant-seq by an in-house R-package ‘McRIPtools’, which tests for enrichment using a binomial-distribution-based model. This procedure resulted in 4,178 enriched peaks called in at least two out of the three biological replicates, with false discovery rate (FDR) < 0.05; 2,185 of these showed at least an average enrichment of 2-fold (FCmean > 2, where FC = CountIP-seq/Countquant-seq) across biological triplicates.

Splice-in analysis and calibration curve for mA stoichiometry estimation. Sequencing data were aligned to the in-silico RNA sequences after deduplication and adaptor cut. Reads were counted by bam-readcount. The observed mutation rates at different mA 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 mA. We used a fraction parameter (fA) to account for the fraction of read-through cDNA that came from mA, and assumed an averaged mutation rate of 100% mA as R0. The observed mutation rate y (as a percentage) and mA fraction x (as a percentage) can be expressed by the following equation,

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

where B is the background mutation rate (that is, the mutation rate at A); 100 − x represents the fraction of A; and 1 − fA accounts for the 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 mA site.

We fit the spike-in mutation rate and mA fraction using the above equation with the non-linear curve fitting by the Prism software. The best-fit values from the data resulted in B0 = 0.32 and fA = 0.22, which were used in the calibration equation (Fig. 5a) for the estimation of mA stoichiometry in mA-quant-seq.

Motif enrichment and metagene analyses. Nine-mer sequences with 4 nucleotides before and after the identified mA sites in mA-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. The metagene profile was plotted with the ‘Guitar’ R package46.

Statistical tests. Data are presented as mean with s.e.m or s.d. as noted in each case. All n values 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 mA-IP-seq and mA-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 mA oligonucleotide library, mA-IP-seq and mA-quant-seq are available in the Supplementary Data.

References
38. Le Grice, S. F. J., Cameron, C. E. & Benkovic, S. J. in Methods in Enzymology Vol. 262 (ed. Campbell, J.) 130–144 (Acad. Press, 1995).
39. Lee, T. S. et al. BglBrick vectors and datasheets: A synthetic biology platform for gene expression. J. Biol. Eng. 5, 12 (2011).
40. Mishina, Y., Chen, L. X. & He, C. Preparation and characterization of the native iron(II)-containing dna repair AlkB protein directly from Escherichia coli. J. Am. Chem. Soc. 126, 16930–16936 (2004).
41. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBNet. J. 2011, 3 (2011).
42. Bushnell, B. BRMap Package (SourceForge, 2019); https://sourceforge.net/projects/bbmap/
43. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
44. 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).
45. Lawrence, M. et al. Software for computing and annotating genomic ranges. PLOS Comput. Biol. 9, e1003118 (2013).
46. Lin, Y.­C. et al. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. Nat. Commun. 5, 4767 (2014).
47. Chen, L. Characterization and comparison of human nuclear and cytosolic editomes. Proc. Nail Acad. Sci. USA 110, E2741 (2013).
48. Rigby, R. A. & Stasinopoulos, D. M. Generalized additive models for location, scale and shape. J. R. Stat. Soc. Ser. C. (Appl. Stat.) 54, 507–554 (2005).
49. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
50. Bailey, T. L. et al. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–W208 (2009).
51. Cui, X. et al. Guitar: An R/Bioconductor package for gene annotation guided transcripomic analysis of RNA-related genomic features. BioMed. Res. Int. 2016, 8 (2016).
Reporting Summary

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

Statistical parameters

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

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection: Fluorescence was collected and processed on the BioTek Synergy Neo2 multi-mode reader; electrophoresis images are collected by UVP GelDoc-it Imaging System with LMS-26 Transilluminator and Bio-Rad ChemiDoc MP Imaging System.

Data analysis: Data was averaged and presented with Microsoft Excel for Mac (version 16) and R Studio (version 1.1.453); electrophoresis images were quantified with ImageJ (version 2.0.0); sequencing data were visualized using the Integrative Genomics Viewer (Version 2.4.16).

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

Data

Policy information about availability of data

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

- Raw and processed m1A-IP-seq and m1A-quant-seq data are available at NCBI Gene Expression Omnibus, accession number GSE123365. The data that support the findings of this study are available from the corresponding author upon request.
**Field-specific reporting**

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

- [x] Life sciences
- [ ] Behavioural & social sciences
- [ ] Ecological, evolutionary & environmental sciences

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

---

**Life sciences study design**

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

| Sample size | No statistical methods were used to determine sample size. Biochemical experiments were performed to achieve consistent results and allow statistical analyses. Triplicate seq experiments were performed. State-of-the-art is currently 1-2 replicates and triplicate is an advance and will help lessen false-positives. |
| --- | --- |
| Data exclusions | No data were excluded from analyses. |
| Replication | We completed all in vitro enzymatic assays multiple separate times with at least two biological replicates. |
| Randomization | Biological triplicates for m1A-IP-seq and m1A-quant-seq were sequenced together. |
| Blinding | Blinding is not relevant in this study because no participants are involved; this study only involves basic molecular experiments. |

---

**Reporting for specific materials, systems and methods**

### Materials & experimental systems

| Involved in the study |
| --- |
| [ ] Unique biological materials |
| [ ] Antibodies |
| [ ] Eukaryotic cell lines |
| [ ] Palaeontology |
| [ ] Animals and other organisms |
| [x] Human research participants |

### Methods

| Involved in the study |
| --- |
| [ ] ChIP-seq |
| [ ] Flow cytometry |
| [x] MRI-based neuroimaging |

---

**Unique biological materials**

Policy information about availability of materials

**Obtaining unique materials**

Engineered reverse transcriptases (DNA and pure proteins) are available from the corresponding authors upon request.

**Eukaryotic cell lines**

Policy information about cell lines

**Cell line source(s)**

HEK293T (ATCC)

**Authentication**

HEK293T cells were recently purchased directly from ATCC and maintained at low passage.

**Mycoplasma contamination**

HEK293T cells in use showed no detectable mycoplasma infection.

**Commonly misidentified lines**

HEK293T cells used because they were used in previous m1A mapping studies.

---

**Magnetic resonance imaging**

**Experimental design**

| Design type |
| --- |
| n/a |
### Acquisition

| Parameter                      | Details                                           |
|-------------------------------|--------------------------------------------------|
| Design specifications         | n/a                                              |
| Behavioral performance measures | n/a                                             |
| Imaging type(s)               | n/a                                              |
| Field strength                | n/a                                              |
| Sequence & imaging parameters | Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle. |
| Area of acquisition           | State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined. |
| Diffusion MRI                 | □ Not used                                       |

### Preprocessing

| Process                      | Details                                                                 |
|------------------------------|-------------------------------------------------------------------------|
| Preprocessing software       | Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.). |
| Normalization                | If data were normalized/standardized, describe the approach(es); specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization. |
| Normalization template       | Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized. |
| Noise and artifact removal   | Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration). |
| Volume censoring             | Define your software and/or method and criteria for volume censoring, and state the extent of such censoring. |

### Statistical modeling & inference

| Type                          | Details                                                                 |
|-------------------------------|-------------------------------------------------------------------------|
| Model type and settings       | Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation). |
| Effect(s) tested              | Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used. |
| Specify type of analysis:     | □ Whole brain □ ROI-based □ Both                                        |
| Statistic type for inference  | Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods. |
| Correction                    | Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo). |

### Models & analysis

| Analysis Type                  | Details                                                                 |
|-------------------------------|-------------------------------------------------------------------------|
| n/a                           | Involved in the study on functional and/or effective connectivity      |
| □ Functional and/or effective connectivity |                                                     |
| □ Graph analysis               |                                                                        |
| □ Multivariate modeling or predictive analysis |                                                            |