Sequencing and structure probing of long RNAs using MarathonRT: a next-generation reverse transcriptase

Li-Tao Guo\textsuperscript{a,1}, Rebecca L. Adams\textsuperscript{a,1,*}, Han Wan\textsuperscript{a}, Nicholas C. Huston\textsuperscript{a}, Olga Potapova\textsuperscript{a}, Sara Olson\textsuperscript{d}, Christian M. Gallardo\textsuperscript{e}, Brenton R. Graveley\textsuperscript{d}, Bruce E. Torbett\textsuperscript{e}, Anna Marie Pyle\textsuperscript{a,b,c}

\textsuperscript{a}Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA

\textsuperscript{b}Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA

\textsuperscript{c}Department of Chemistry, Yale University, New Haven, Connecticut 06520, USA

\textsuperscript{d}Department of Genetics and Genome Sciences, Institute for Systems Genomics, UConn Health, Farmington, Connecticut 06030-6403, USA

\textsuperscript{e}The Scripps Research Institute, La Jolla, CA 92037, USA

Abstract

Reverse transcriptase (RT) enzymes are indispensable tools for interrogating diverse aspects of RNA metabolism and transcriptome composition. Due to the growing interest in sequence and structural complexity of long RNA molecules, processive RT enzymes are now required for preserving linkage and information content in mixed populations of transcripts, and the low-processivity RT enzymes that are commercially available cannot meet this need. MarathonRT is encoded within a eubacterial group II intron and it has been shown to efficiently copy highly structured long RNA molecules in a single pass. In this work, we systematically characterize MarathonRT as a tool enzyme and optimize its performance in a variety of applications that include single cycle reverse-transcription of long RNAs, in-cell SHAPE-MaP using ultra-long amplicons and the detection of natural RNA base modifications. By diversifying MarathonRT reaction protocols, we provide an upgraded suite of tools for cutting-edge RNA research and clinical application.
INTRODUCTION

The ability to interrogate the full spectrum of RNA molecules within a cellular compartment is essential for understanding transcriptome dynamics and its impact on gene expression. Accurate analysis of transcript populations is fundamentally dependent on the conversion of RNA to complementary DNA (cDNA), which is made possible by reverse-transcriptase enzymes (RTs). Modern RNA technologies, such as RT-PCR and RNA-Seq, were developed using retroviral RTs, and they have been optimized around the inherent shortcomings of that enzyme class. Even highly optimized retroviral RTs suffer from low processivity, which results in mostly truncated cDNA products that arise from prematurely terminated primer extension. Conventional RT enzymes are unable to copy long RNA molecules in a single pass, so the sequences of long RNA molecules, such as viral genomes, long noncoding RNAs (lncRNAs), or mRNA transcripts are not directly determined; rather they are inferred by assembling short reads on a reference genome. As a result, one can only report an average sequence, not physically linked genes, and cannot differentiate the sequences of RNAs within a complex mixture, such as alternative splice isoforms or the spectrum of mutations within a viral swarm. As a result, we lose vital information on dynamic processes important for analyzing gene expression. Furthermore, due to the low cDNA yields of retroviral RTs, the outlook is even worse for RNA species that are low in abundance. An ideal RT would preserve the information transfer between RNA and DNA, regardless of RNA template length or structural content, thereby facilitating a diversity of downstream applications. MarathonRT, a recently-discovered group II intron-encoded RT\cite{1} is an enzyme with properties that closely approach the characteristics of an ideal RT\cite{2}.

In addition to their utility for analyzing primary RNA sequence, RT enzymes enable accurate RNA secondary structure determination by sensing the modifications deposited by structure-specific chemical probes, such as 2′-hydroxyl acylation reagents (SHAPE-MaP)\cite{3} and dimethyl sulfate (DMS-MaP)\cite{4}. In SHAPE-MaP experiments, 2′-hydroxyl-selective reagents are reacted with an RNA of interest, forming covalent 2′-O-adducts at conformationally flexible RNA nucleotides. Acylated nucleotides are sensed by the RT and misread, and the SHAPE adducts are recorded as elevated mutational signals in next-generation sequencing libraries. The current SHAPE-MaP workflow relies on SuperScript II (SSII), which is an RT derived from the Moloney Murine Leukemia Virus (MMLV) RT, to detect SHAPE-modified nucleotides. Due to its intrinsically low processivity, SSII can copy only a few hundred nucleotides along a modified RNA template, which makes the extraction of structural information from long modified RNA extremely laborious. Furthermore, structural information for low-abundance target RNAs is often unattainable due to low cDNA yields. As such, the limitations imposed by currently available commercial RTs make it very difficult to perform structure mapping on long, or even moderate-length RNAs of high biological interest.
In this study, we systematically characterized and optimized the conditions for utilizing MarathonRT, resulting in high performance on-demand RT applications including standard reverse transcription as well as in vitro and in vivo SHAPE-MaP. The ultra-high processivity of MarathonRT results in abundant production of cDNA products of at least 2500 nt on modified templates, which far exceeds the capabilities of SSII. In the process of this investigation, we also discovered that MarathonRT can directly detect natural modifications within cellular RNA species and we create a finely-tuned set of reaction conditions for high-throughput profiling of novel RNA targets. When used under optimized conditions, MarathonRT brings out the full potential of reverse-transcription in RNA science and technology.

RESULTS

Establishing the specific activity assay

In order to ensure reproducibility, it was important to establish standardized metrics for performance of the enzyme. To this end, we defined one unit of MarathonRT as the amount of enzyme that can incorporate 1 nmole of dTTP in 30 minutes at 42°C when using poly(rA)•oligo(dT)18 as the substrate. A representative stock solution of MarathonRT at 5 μM contains 20 ± 1 units/μL, according to measurements from four independently-purified batches of enzyme (Suppl. Table 1).

Strategies for buffer optimization

To optimize primer utilization efficiency and overall performance of the enzyme, we undertook a comprehensive analysis of reaction conditions and optimized the protocols for utilization of MarathonRT. Ionic strength and pH have a major influence on stability and substrate affinity of nucleic acid enzymes\[5\], and we therefore optimized the influence of these parameters on primer extension by MarathonRT. Given that reverse transcription by MarathonRT is performed at 42°C, above the optimal growth temperature of 37°C of its host organism, Eubacterium rectale, we also examined several classes of protein stabilizing agents.

The optimization process was sequentially implemented. Using Domain 3 from lncRNA RepA as an RNA template (RepA D3, 622 nt), and our previously-developed reaction buffer as the initial buffer\[2\], we first explored the effect of pH on MarathonRT. To quantify MarathonRT efficiency, a “primer extension efficiency” metric was calculated as the percentage of total extended primers relative to total primers. A range of pH from 7.5 to 8.5 was tested using K-HEPES as used in the initial buffer (Suppl. Fig. 1A). Primer extension was highest at pH 8.3. This observation was reproduced with K-TAPS (Suppl. Fig. 1B) and Tris-HCl (Fig. 1A), confirming the pH optimum of 8.3. Since primer extension was most efficient in Tris-HCl (41% efficient at pH 8.3), we continued optimization using this buffer. We then titrated the concentration of MgCl2 and confirmed that 2 mM Mg2+ is optimal for MarathonRT at pH 8.3 (Fig. 1B and Suppl. Fig. 2). To explore the effect of ionic strength, we then varied the concentration of KCl over a range of 50 mM to 300 mM, observing that primer extension peaked at ~52% in 200 mM KCl (Fig. 1C). A similar trend was observed when replacing KCl with NaCl (Suppl. Fig. 3A) or NH4Cl (Suppl. Fig. 3B), but KCl at 200
mM was selected for further optimization due to its physiological relevance and support of high enzyme efficiency.

During the second phase of the analysis, we tested the protein stabilizing agents betaine, trehalose, glycerol, bovine serum albumin (BSA) and PEG 8000 (Fig. 1D and Suppl. Fig. 4). Betaine and trehalose are common reverse transcription additives that are known to increase cDNA yield\cite{6}. We observed that 1 M betaine and 0.6 M trehalose moderately increased primer extension efficiency to 58% and 63%, respectively, though the combination of the two did not result in an additive effect. Glycerol is known to stabilize proteins and is widely used for long-term storage of purified proteins, although it is rarely added to specifically boost enzyme activity. Surprisingly, in the presence of 20% glycerol, primer extension efficiency was dramatically increased to ~89% (Fig. 1D), and nearly all of the primer was extended to full length by MarathonRT (Fig. 1F). BSA and PEG 8000 also serve as molecular crowding agents, but they were not observed to substantially improve cDNA yield (Suppl. Fig. 4).

Taken together, we have identified an optimized buffer for MarathonRT that contains 50 mM Tris-HCl pH 8.3, 200 mM KCl, 2 mM MgCl$_2$, 5 mM DTT and 20% glycerol. We confirmed that improved cDNA synthesis in this buffer is observed with other RNA templates, including the a$\alpha$y group IIB intron (931 nt), the 5' end of hepatitis C virus (HCV) genome (1271 nt), Domain 1 of HOTAIR (HOTAIR D1, 532 nt) and the Oceanobacillus iheyensis group IIC intron (803 nt), which are RNA molecules that vary significantly in length, structure, and G-C content. For all of these templates, efficiency was improved 1.3 – 2.4 times in the optimized buffer (Suppl. Fig. 5). Importantly, MarathonRT’s high processivity is retained in the optimized buffer, as indicated by the high yield of full-length cDNA products (Fig. 1F and Suppl. Fig. 5A).

**Evaluation of MarathonRT performance under suboptimal and inhibitory conditions**

Because reverse transcriptases are employed in a diversity of biotechnological applications, it was important to assess whether MarathonRT can maintain robust cDNA yields under suboptimal buffer conditions that would be encountered with downstream applications. Specifically, we examined primer extension under common suboptimal conditions of pH, Tris-HCl pH 7.5 (Fig. 1E and 1F), and in a common suboptimal buffer that contains Na-HEPES pH 8.3 (Suppl. Fig. 6). For the conditions tested, we replaced the optimal KCl salt with suboptimal NaCl to further challenge the activity of MarathonRT. When any of the three effective stabilizing agents, betaine, trehalose or glycerol, were added to the two suboptimal buffers, a marked enhancement of primer extension efficiency was observed. Specifically, robust activity was almost completely maintained as long as 20% glycerol was added to the two suboptimal buffers (Fig. 1E and 1F and Suppl. Fig. 6). Based on these findings, we conclude that MarathonRT tolerates HEPES buffer, and retains efficient function over a range of pH values (from 7.5 to 8.3) and salts, indicating a broad spectrum of possible applications for MarathonRT.

Another factor that would be expected to impact the function of MarathonRT is the presence of RT inhibitors that are commonly used during RNA isolation or present in the biological samples. These include reaction conditions previously shown to inhibit PCR\cite{7}. We explored
the effects of several potential RT inhibitors on the efficiency of reverse transcription by MarathonRT in the optimized reaction buffer. Specifically, we tested the effects of BugBuster, phenol, chloroform, guanidinium chloride, ethanol, isopropanol, formaldehyde, bile salt and humic acid. We calculated primer efficiency at the indicated final concentration of inhibitor and normalized this value against conditions without an inhibitor. In general, MarathonRT displayed only modest, if any, decrease in activity in the presence of most inhibitors tested in this study, except for phenol and isopropanol (Suppl. Fig. 7). Therefore, removal of residual phenol from RNA samples is required for subsequent reverse transcription, and RNA samples precipitated with isopropanol should be thoroughly dried before being re-dissolved in water. Taken together, however, the data indicate that MarathonRT is highly resistant to a diversity of chemical inhibitors.

**MarathonRT activity and stability are compromised at increased temperatures**

To explicitly evaluate how MarathonRT activity is affected at higher temperatures, we performed reverse transcription with MarathonRT at 47°C and 50°C in several buffer conditions (Fig. 2A). The buffers used here include the initial buffer, the optimized buffer without additive, and the optimized buffer supplemented with betaine, trehalose, or glycerol. Primer extension was performed using the RepA D3 template and compared to activity at 42°C. Primer extension efficiency dropped to less than 20% in the absence of a protein stabilizing agent at 47°C, with addition of betaine, trehalose, or glycerol partially restoring activity to >40%. At 50°C, cDNA product yield was further decreased with or without additives, making this temperature unusable for MarathonRT.

To quantitatively evaluate MarathonRT thermostability, we performed fluorescence thermal shift assays using GloMelt™, a buffer-insensitive fluorescent dye (Biotium)\[8\]. Melting temperature was calculated as the midpoint value of the slope in the fluorescence intensity values as a function of temperature. The melting temperatures of MarathonRT (in the absence of template) were found to be 33.2°C in the initial buffer, 35.2°C in the optimized buffer, 33.3°C in the optimized buffer with betaine, 33.7°C in the optimized buffer with trehalose, 36.3°C in the optimized buffer with both betaine and trehalose, and 40.2°C in the optimized buffer with glycerol (Fig. 2B). Increases in the melting temperature of MarathonRT in specific improved buffers (which contain protein stabilizing agents) track to some extent with improvements in the efficiency of primer extension. However, betaine and trehalose did not demonstrate an apparent stabilizing effect despite their improvement in primer extension (Fig. 2A), which may be due to a stabilizing influence of the template-primer substrate that could not be measured in the thermal shift assay (nucleic acid disrupts the signal from the dye and therefore we could only measure stability of the apo-enzyme). Overall, the thermal shift assay reinforced the selection of glycerol as an optimal additive for MarathonRT and confirmed that reaction temperatures above 42°C are not recommended for the wild-type enzyme.

**Detection of RNA modifications and improved RNA structure prediction by MarathonRT**

An important application for RT enzymes is the detection of RNA adducts that are introduced by chemical probes during RNA structure-probing experiments. A variety of structure-sensitive chemical probes have been developed, including DMS, which
preferentially methylates the Watson-Crick face of unpaired A and C nucleotides, and SHAPE reagents, which acylate the 2′-OH of flexible nucleotides. It was previously shown that a thermostable group II intron RT (TGIRT) primarily introduces RT stops at DMS-modified positions in buffers that contains magnesium\[9\]. In this way, DMS-MaPseq has been used to perform genome-wide and targeted RNA structure probing \textit{in vivo}\[4\]. Similarly, MarathonRT favors reverse transcription stops rather than mutation incorporation in the presence of magnesium, an attribute that is useful for detection of natural nucleotide modifications (\textit{vide infra}).

In a more sensitive method of chemical probing (Probe-MaP)\[10\], the adducts induced by chemical probes are detected not from the positions of RT stops, but rather from the tendency of an RT enzyme to introduce mutations at the positions of chemically-modified nucleotides (increased mutation rates), which are then detected using RNA sequencing. To induce nucleotide mis-incorporation by SSII at SHAPE-modified nucleotides, 6 mM manganese is routinely used in place of magnesium\[3\], as Mn$^{2+}$ has long been known to relax the specificity of DNA polymerases\[11\]. We therefore wondered whether MarathonRT could be induced to incorporate mutations at probed nucleotides in the presence of manganese.

To evaluate whether MarathonRT can be implemented as a tool for Probe-MaP experiments, we performed primer extension experiments over a range of Mn$^{2+}$ concentrations using DMS-modified HCV genomic RNA. This template was chosen because it is highly structured and it contains a well-characterized internal ribosome entry site (IRES) that has been characterized by chemical probing\[3\] and even by high resolution cryo-EM structure determination\[12\] (Suppl. Figure 8A). We observed that the best Mn$^{2+}$ concentration range for full-length product generation with MarathonRT is much lower than conventional applications, and that the optimum is between 0.5–2.0 mM, whether the template is modified or not (Fig. 3A, lanes 4–6, 12–14). This condition results in a dramatic improvement in the yield of full-length cDNA products compared to standard SHAPE-MaP conditions with SSII (Fig. 3A, lanes 2 & 10).

To analyze whether MarathonRT incorporates mutations at 2′-acylated (SHAPE-modified) nucleotides, we folded the \textit{in vitro} transcribed HCV genome\[10\] and then modified the RNA with a variety of chemical probes including the SHAPE reagents 1M7, NAI, and FAI, together with DMS. We then generated high-throughput sequencing libraries from modified HCV IRES RT products\[10\] using both the standard MaP condition (SSII, 6 mM Mn$^{2+}$) and MarathonRT in the presence of 1 mM Mg$^{2+}$, 1 mM Mn$^{2+}$, and 6 mM Mn$^{2+}$. We found that, in the presence of 1 mM Mn$^{2+}$, MarathonRT displayed higher mutation rates using libraries prepared from modified RNA templates (Suppl. Fig. 8B, bars 8–10,12) relative to unmodified RNA templates (Suppl. Fig. 8B, bars 7 & 11), indicating that MarathonRT can incorporate mutations at probed nucleotides, rather than merely stopping. Notably, the background MarathonRT mutation rate for unmodified RNA is approximately 10-fold lower when libraries are prepared with 1 mM Mn$^{2+}$ relative to the standard MaP conditions (6 mM Mn$^{2+}$) (Suppl. Fig. 8B, bars 11 vs. 13). This is because Mn$^{2+}$ is inherently mutagenic for polymerase enzymes, and there is a dose-dependence to levels of nucleotide misincorporation\[11\]. This drastic reduction in background misincorporation improves the
overall signal to noise ratio. For example, when MarathonRT reverse transcribes in the presence of 6 mM Mn\textsuperscript{2+}, the distribution of mutation rates for (+) and (−) DMS conditions are highly overlapping (Suppl. Fig. 8B, bars 13 vs. 14) whereas in the presence of 1 mM Mn\textsuperscript{2+}, the distributions are well separated (Suppl. Fig. 8B, bars 11 vs. 12).

Because DMS and SHAPE reagents modify single-stranded nucleotides, it would be expected that MarathonRT introduces more mutations at single-stranded regions. As the HCV IRES has been crystallized, we were able to unambiguously evaluate whether the observed increase in mutation rates from modified templates are derived from single- or double-stranded nucleotides. Consistent with the model, mutation rates predominantly increased at single-stranded nucleotides when 1M7, NAI, FAI, and DMS were used to modify the HCV IRES template (Fig. 3B, bars 5–16). We observed a higher mutation rate for DMS modified RNA relative to unmodified RNA in the presence of Mg\textsuperscript{2+}, as reported for TGIRT\textsuperscript{[4]} (Fig. 3B, bars 1–4 and Suppl. Fig. 8, bars 5 & 6), but there are some high reactivity nucleotides that map to double-stranded regions of the HIV IRES (Suppl. Fig. 9A). Importantly, when 1 mM Mn\textsuperscript{2+} is used, reactivities calculated from SHAPE- and DMS-MaP by MarathonRT accurately match the HCV IRES structure (Suppl. Fig. 9B).

A major benefit of SHAPE-MaP is an improved accuracy for structure prediction. As the HCV IRES is known to contain a pseudoknot, a minimum-free energy optimization algorithm that allows for the prediction of pseudoknots was required for our secondary structure prediction. To this end, we used ShapeKnots for calculated structure predictions, with chemical reactivities determined using ShappeMapper2 input as pseudo-free energy constraints. We found that reactivities derived using 1M7, NAI, and FAI resulted in almost completely accurate prediction over a range of slope and intercept values (Suppl. Fig. 10). Note that MarathonRT has a higher mutation rate for NAI- and FAI-treated RNA than 1M7-treated RNA (Fig. 3B and Suppl. Fig. 8B), but 1M7-treated RNA resulted in reactivities that accurately constrained structure prediction, albeit over a narrower slope and intercept range than for NAI and FAI (Suppl. Fig. 10).

As another determinant for structure-prediction accuracy, we analyzed the types of mutations introduced by MarathonRT in the presence of 1mM Mn\textsuperscript{2+}. We found that the majority of mutations introduced by MarathonRT are base substitutions (mismatches), regardless of the SHAPE probe that is used (Fig. 4A). This is in sharp contrast with the mutations that are introduced by SSII, which induces a far more complex mutational profile that includes deletions, insertions, multinucleotide mismatches, and complex mutations at positions of RNA modification\textsuperscript{[13]} (Fig. 4B). Analyzing the mutational profile of MarathonRT and SSII on DMS-modified templates in depth reveals that most mutational substitutions occur at adenosine and cytidine nucleotides, which is expected because DMS preferentially modifies imino heteroatoms on the base pairing face of A and C (Fig. 4C and 4D, colored slices). Importantly, Marathon RT copies 5% of DMS adducts as complex mutations other than base substitutions, whereas SSII encodes 29% of DMS base-adducts as more complex mutational types (Fig. 4C and 4D, grey slices). According to Busan and Weeks\textsuperscript{[14]}, simple substitutional mismatches do not generate ambiguity during sequencing read alignment, resulting in higher prediction accuracy. This contributes to the high
prediction accuracies observed when SHAPE-MaP is conducted using MarathonRT (Suppl. Table 2).

RNA structure prediction from long amplicons using in-cell SHAPE-MaP

In the native cellular environment, the sensitivity of nucleotides to chemical probes not only reflects the features of RNA structure, but it can also reflect interactions with proteins and other molecules. We therefore set out to examine whether the ultra-high processivity of MarathonRT can be exploited to probe RNA structure and environment in vivo. Our in vitro experiment shows that MarathonRT is more efficient than SSII in generating full-length cDNA product from modified RNA templates in its optimized SHAPE-MaP buffer (Fig. 3A), suggesting that MarathonRT might produce high quality long cDNA products from in vivo modified RNA. This would represent a major unmet need in biotechnology.

Similar to the in vitro experiment, we used the well-characterized HCV IRES structure as the benchmark for evaluating accuracy of in-cell SHAPE-MaP with MarathonRT. We used both a short amplicon (397 nt, the HCV IRES region) and a long amplicon (2521 nt, with HCV IRES located at the end of the cDNA product) to examine the consistency of in-cell SHAPE-MaP performance (Fig. 5A). Meanwhile, standard in-cell SHAPE-MaP with SSII was also performed as a control experiment using the same template. The experimental procedure was as described in the Methods. As shown in Suppl. Fig. 11, SSII is unable to generate detectable long amplicons on both modified and unmodified template after 35 cycles of PCR amplification. Therefore, only the cDNA product from short amplicon by SSII was used to construct a high-throughput sequencing library for comparison with MarathonRT.

For our in vivo experiments with MarathonRT, the conditions we previously optimized for in vitro SHAPE-MaP (1mM Mn²⁺, pH 8.3) did not yield reproducible results, but a small reduction in pH was sufficient to obtain robust conditions for in vivo applications (1 mM Mn²⁺, pH 7.5) (Fig. 5B, short: bars 5 vs. 7, long: 9 vs. 11). Using reactivities calculated using ShapeMapper2, we determined that data collected with both short and long amplicons mapped well to the HCV IRES structure, with higher reactivities at single-stranded nucleotides as expected (Fig. 5C). The reactivities were then used to predict the HCV IRES pseudoknot structure de novo over a range of slope and intercept values (Suppl. Fig. 12). Collectively, our results show that MarathonRT can be used in Probe-MaP experiments at low Mn²⁺ and that its high processivity can be leveraged to greatly simplify SHAPE-MaP library construction, producing high-quality data that can be used to accurately predict RNA secondary structure.

MarathonRT incorporates mutations at natural in vivo RNA modification sites

During our studies of in vivo structure-probing with MarathonRT, we were surprised to observe that, in the presence of 1 mM Mn²⁺, MarathonRT cDNA products displayed a higher mutation rate when generated from HCV RNA isolated from cells versus HCV RNA generated by in vitro transcription (Suppl. Fig. 13). A similar difference between in vitro and in vivo mutation rates was not observed for SSII. We hypothesized that the higher in vivo mutation rate for MarathonRT might be due to the incorporation of mutations at naturally
post-transcriptionally modified nucleotides. Because ribosomal RNA (rRNA) is among the most homogenously modified RNA in cells, containing well-established sites for a diversity of nucleotide modifications, we reasoned that 18s and 28s rRNAs would be ideal RNA templates for evaluating the ability of MarathonRT to detect natural modifications. We focused our initial analysis on rRNA nucleotides known to harbor 2'-O-methylation marks for two reasons. First, natural 2'-OMe groups resemble the adducts deposited on the 2'OH group by electrophilic SHAPE compounds, suggesting that 2'-OMe modifications might be readily detected by MarathonRT. Secondly, 2'-OMe groups are the most abundant natural modification in rRNAs, ensuring that any signal we might observe carries statistical weight.

After analyzing the MarathonRT extension data on 18s and 28s rRNAs, we observed strong mutational incorporation at all known positions of natural 2'-OMe modification. This effect was observed in a broad range of metal ion conditions, including the original Mg\textsuperscript{2+} buffer. Even stronger misincorporation is seen over a range of increasing Mn\textsuperscript{2+} conditions (Fig. 6A). The fact that one can observe robust 2'-OMe detection, but not 2'-acyl group detection in magnesium conditions (Supp. Fig 8, bar 1–4) suggests that MarathonRT is more sensitive to 2'-OMe groups than the 2'-acyl groups introduced by SHAPE reagents.

Pseudouridine is the second most abundant modification of rRNAs, and we therefore sought to determine if they are detectable by MarathonRT. Like 2'-O-methylated nucleotides, bulk populations of pseudouridylated nucleotides display higher mutation rates than unmodified uridines, suggesting that MarathonRT is able to detect these natural modifications as well (Fig 6B). Indeed, in the presence of ≥1mM Mn\textsuperscript{2+}, a significant increase in mutation rate was observed for pseudouridylated nucleotides relative to unmodified uridines (Fig 6B), thereby establishing that MarathonRT is capable of detecting natural pseudouridine nucleotides. After optimization, we determined that an RT buffer containing 2mM Mn\textsuperscript{2+} is ideal for detection of natural modifications due to the robust mutational signal observed for at least the two most abundant types natural modifications (both 2'-OMe and pseudouridine).

Importantly, we avoid the apparent inhibitory effect of high Mn\textsuperscript{2+} on MarathonRT cDNA yields that is encountered when using Mn\textsuperscript{2+} concentrations >2mM (Fig. 3A).

While our data indicate that 2'-O-methylated and pseudouridylated nucleotides can be detected by MarathonRT, the analysis described above is appropriate only on RNA molecules that contain known positions and types of modification (example, Fig 6 A and B). In practice, on an RNA of unknown composition, one would reverse-transcribe an RNA of cellular origin in either standard RT conditions (i.e., 1mM Mg\textsuperscript{2+}) or in the presence of 2mM Mn\textsuperscript{2+}, construct next-generation sequencing libraries, and then identify potentially modified nucleotides on the basis of a signal that differentiates them from all other nucleotides. In the simplest case, mutation rates alone would be sufficient to identify bases that potentially harbor natural modifications. As is evident in Fig 6C and 6D (18s and 28s rRNA respectively), an x-y scatter plot of mutation rates calculated for cDNA generated by MarathonRT in the presence of standard RT buffer or 2mM Mn\textsuperscript{2+}, 2'-OMe nucleotides (yellow) cluster separately from unmodified nucleotides (grey). This suggests that 2'-OMe modifications can in fact be identified on the basis of their mutation rates in the two RT buffers alone. Importantly, this pattern of 2'-OMe clustering holds for both 18S and 28S rRNAs in multiple biological replicates (Suppl. Fig. 14). For an RNA of unknown natural...
modification status, our data suggests that nucleotides clustering within this quadrant are likely to harbor specific 2’-OMe modifications, warranting experimental follow-up.

In contrast, our analysis revealed that pseudouridylated nucleotides (blue) do not separate strongly from unmodified nucleotides (grey) on the basis of their mutation rates alone (Fig 6C and 6D), suggesting additional pieces of data would be required to identify putative pseudouridines. Indeed, a further dissection of mutation types induced by MarathonRT at pseudouridylated sites reveals that the rate of U-to-A substitution was significantly and reproducibly increased on both 18s and 28s rRNA templates relative to unmodified uridines, whereas the U-to-G and U-to-C substitution rates were not significantly different (Suppl. Fig. 15). This may represent a signature for pseudouridylation detection by MarathonRT. We anticipate that this type of unique mutational pattern, in addition to an elevated mutation rate in 2mM Mn\(^{2+}\) might be useful for detection of novel pseudouridylation sites, but would require significantly more optimization both experimentally and computationally.

In addition to abundant 2’-OMe and pseudouridylated nucleotides, 18s and 28s rRNAs harbor a diversity of additional rare modifications. Statistical analysis of our data suggests that MarathonRT is able to detect some of these species. For example, in 18s rRNA, positions containing 1-methyl-3-(3-amino-3-carboxypropyl)-pseudouridine (m\(^1\)acp\(^3\)ψ) displayed the highest mutation rate of any nucleotide in the presence of both Mg\(^{2+}\) and Mn\(^{2+}\), over multiple replicates (Fig. 6C, Suppl. Fig. 14). Additionally, the 7-methylguanosine (m\(^7\)G) modified nucleotide on 18S rRNA displayed a reproducible pattern of mutation rates that is similar to that of the 2’OH-methylated nucleotides (Fig. 6C, Suppl. Fig. 14). In the 28s rRNA, both 3methyluridine (m\(^3\)U) and 1-methyladenosine (m\(^1\)A) modified nucleotides also appear as the nucleotides with exceptionally high mutational rates under both Mg\(^{2+}\) and Mn\(^{2+}\) conditions (Fig. 6D and Suppl. Fig. 14). However, further work is needed to compile descriptive, statistically significant mutational profiles for these rare modifications.

**DISCUSSION**

In this study, we systematically investigated the reaction conditions and structure-probing applications for the ultra-processive reverse-transcriptase MarathonRT. Our experiments show that MarathonRT is a robust enzyme that retains nearly full activity under a variety of conditions, including those that are normally detrimental to polymerase activity. For example, MarathonRT retained full activity at 10 ng/μL of humic acid in contrast to Taq DNA polymerase that is inactivated by 4 ng/μL humic acid\(^{15}\) (Suppl. Fig. 7). Importantly, the adaptability of MarathonRT to various reaction conditions enabled us to optimize reverse transcription for both in vitro and in-cell SHAPE-MaP as well as the detection of natural modifications, using both short and long amplicons.

Although MarathonRT retains activity in diverse conditions, thermal shift experiments show that MarathonRT is not an exceptionally thermostable enzyme. However, the ability of MarathonRT to generate full-length cDNA products at relatively low temperatures (i.e. 42°C) is actually an advantage over other processive enzymes, such as TGIRT, which require reaction temperatures as high as 60°C to resolve structures in the RNA template\(^{16}\). High
temperature conditions, particularly in combination with divalent cations, are disadvantageous because they cause hydrolytic breaks in the RNA template, thereby undercutting any processive capabilities of the enzyme. We suspect that the enhanced cDNA yield by MarathonRT may be partly attributable to its ability to reverse transcribe even highly structured templates at lower temperatures, conditions in which the RNA template is more stable. While the thermal shift assay indicates that apo-MarathonRT has a relatively low melting temperature, binding of template-primer is expected to greatly stabilize the enzyme and improve overall thermostability, as previously reported for other polymerases\(^5\). It would have been ideal to test this directly, however, inclusion of primer-template in the thermal shift buffer causes fluorescence intensity of GloMelt dye to peak due to interactions with the nucleic acid, making thermal shift assays challenging in the presence of primertemplate. Detailed biophysical analysis of MarathonRT stability properties would be an ideal subject for future studies using other methods.

Of particular importance in this study is the fact that DMS-/SHAPE-MaP experiments can be carried out using MarathonRT. The enzyme is significantly more efficient at generating full-length cDNA products from modified templates when compared to SSII, which is the enzyme conventionally employed for Probe-MaP experiments (Fig. 3A). The exceptional processivity of MarathonRT is reflected in its ability to generate ultra-long cDNA products (2.5 kb) from modified templates, an attribute not shared by SSII (Suppl. Fig. 11). Because SSII cannot reliably transcribe cDNA templates >700nt from modified templates, structure probing of large RNAs requires the use of short, overlapping, tiled amplicons for full coverage. As such, the ability of MarathonRT to generate cDNA templates more than 2.5 kb in length will greatly simplify library construction workflows. For example, the number of amplicons required to probe full-length HCV genomic RNA has been reduced from 12 to 4 when using MarathonRT. It is expected that MarathonRT would be particularly useful for application in RING-MaP (RNA interaction groups by mutational profiling) experiments designed to detect through-space interactions and for profiling diverse RNA conformations of the same molecule. The ultra-high processivity of MarathonRT will increase the distance between detectable through-space interactions, leading to improved resolution of individual RNA conformations and species. However, the use of MarathonRT for RING-MaP applications would require further optimization, including modification conditions that increase the number of co-occurring modifications on a single cDNA molecule with sufficient signal-to-noise ratio, as well as library preparation methods that preserve information about the cDNA molecule from which a modification originated\(^17\).

It has long been known that RNA modifications can alter the behavior of RT enzymes. One case in point is m\(^1\)A, a methylation on the Watson–Crick face of adenosine that, when artificially added in combination of 3-methylcytidine (m\(^3\)C) via DMS modification, can provide a valuable probe of RNA structure\(^4,18\). Such modifications lead either to RT arrest or to misincorporation of a non-complementary dNTP. The ensemble of erroneous events in cDNA synthesis is called a “reverse transcription signature”\(^19\). Some reverse transcription signatures can be induced or enhanced by chemical treatment, allowing a specific modification to be readily distinguished from the background and detected, e.g, by machine-learning-based approaches for evaluating high-throughput sequencing data\(^20\). In this study, we determined that the reverse transcription signatures for 2’-O-methylation and
pseudouridylation generated by MarathonRT can be accurately detected and monitored, resulting in a powerful and simplified new tool for studies of certain post-transcriptional modifications (Fig. 6 and Suppl. Fig. 14). Taking HCV as an example, our preliminary data suggested that HCV genomic RNA may contain natural modifications in vivo (Suppl. Fig. 13), although follow-up studies would be required to identify their positions and chemical identities.

In summary, MarathonRT is a well-folded polymerase that is accurate, ultra-processive, naturally lacking in RNase H activity, and easily purified in a tag-free form due to its high solubility. Systematic optimization of the MarathonRT reaction has resulted in conditions that enable this enzyme to outperform retroviral RTs in every RNA sequencing application tested, while opening the door to new applications for sequencing long RNAs, highly structured RNAs, and RNAs containing post-transcriptional modification. We show that MarathonRT detects natural post-transcriptional modifications and commonly-used chemical probes of RNA secondary structure. Importantly, MarathonRT is readily employed in DMS-/SHAPE-MaP experiments, thereby expanding the types of RNA molecules that can be structurally investigated in vitro and in vivo.

**MATERIALS AND METHODS**

**Protein expression and purification**

Protocols for expression and purification of MarathonRT were adapted from previous studies. *E. coli* Rosetta II (DE3) cells (Millipore) carrying the expression vector pET-6xHis-SUMO-MarathonRT (available at Addgene) were grown to 2.0 to 2.5 of OD600 at 37°C in 2 liters of LB media supplemented with 50 μg/mL kanamycin and 17 μg/mL chloramphenicol. Then the growth temperature was reduced to 16°C followed by addition of IPTG to a final concentration of 0.5 mM to induce the expression of MarathonRT. After 18 hours, cells were harvested by centrifugation at 5,000 rpm for 10 min and stored at −80°C.

For protein purification, cells were resuspended in buffer A (25 mM Na-HEPES, pH 7.5, 1 M NaCl, 10% glycerol and 2 mM β-mercaptoethanol) and lysed by MicroFluidizer at 15,000 psi. Cell debris in the lysate was removed by centrifugation at 13,000 g for 30 min at 4°C. The supernatant from 2 L cell culture was incubated with 2 mL Ni-NTA resin (Qiagen) that was pre-equilibrated with 10 volumes of buffer A at 4°C. After one hour, the cell lysate and Ni-NTA resin were loaded onto a 30-mL gravity column. The packed Ni-NTA resin was washed first with 20 mL of buffer A, then with 20 mL of buffer B (25 mM Na-HEPES, pH 7.5, 500 mM NaCl, 20 mM imidazole, 10% glycerol and 2 mM β-mercaptoethanol) and finally with 20 mL buffer C (25 mM Na-HEPES, pH 7.5, 500 mM NaCl, 30 mM imidazole, 10% glycerol and 2 mM β-mercaptoethanol). Then 20 mL buffer D (25 mM Na-HEPES, pH 7.5, 300 mM NaCl, 300 mM imidazole, 10% glycerol and 2 mM β-mercaptoethanol) was applied to the column to elute the protein. MarathonRT was expressed as a fusion protein in the form of N-His6-SUMO-MarathonRT. To yield the untagged native protein, eluted MarathonRT was incubated with N-His6-Ulp1 protease at 4°C for 1h to cleave off the N-His6-SUMO tag. After protease digestion, the mixture was loaded onto a 5 mL HiTrap SP HP cation exchange chromatography column (GE Healthcare) that had been equilibrated with buffer E (25 mM K-HEPES pH 7.5, 300 mM KCl, 10% glycerol and 1 mM DTT).
protein was eluted with a linear KCl gradient from the loading concentration to 1 M in 20 column volumes (100 mL). The peak fractions were pooled, concentrated to 5 mL and further purified by a HiLoad 16/600 Superdex 200 size exclusion column (GE Healthcare) equilibrated with buffer E. The peak fractions from the Superdex 200 column were checked for purity by SDS-PAGE, pooled, adjusted to 5 μM concentration, flash frozen in liquid nitrogen and stored at −80°C.

**Specific activity assay**

We define one unit of MarathonRT as the amount of enzyme that can incorporate 1 nmole of dTTP in 30 minutes at 42°C when using poly(rA)•oligo(dT)18 as the template substrate. The poly(rA) with an average length of ~500 nt was synthesized by Sigma, and the oligo(dT)18 primer was synthesized by Thermo Fisher.

To obtain the specific activity of MarathonRT, the reverse transcriptase assay was carried out in a 10 μL volume at 42°C. The assay conditions were 50 mM Tris-HCl pH 8.5, 100 mM KCl, 2 mM MgCl₂, 5 mM DTT, 2 μg/μL poly(rA), 0.1 μg/μL oligo(dT)18, 1 mM dTTP, 1 uCi [α−32P]dTTP and 50 nM MarathonRT. After 30 min, the reaction was stopped by adding 2 μL 100 mM EDTA followed by incubation at 95°C for 2 min. Then 1 μL of reaction mixture was spotted onto two 1-cm squares of Hybond-N+ membrane. One of the two squares was washed with 2X SSC buffer three times to remove unincorporated [α−32P]dTTP, and then air-dried. The total radioactivity of the washed Hybond-N+ membrane square represents the incorporated dTTP, and that of the unwashed square represents the total amount of dTTP. To quantify their radioactivity, the two Hybond-N+ membrane squares were exposed to a phosphoimager screen for 2–3 min, which was then scanned with a Typhoon FLA 9500 scanner (GE Healthcare). The image was analyzed and quantified by ImageQuant software (GE Healthcare).

**Reverse transcription assays**

The RNA templates used for RT assays were prepared by *in vitro* transcription, including RepA D3 (residues 998–1630)[21], Group II intron ai5γ with a short 5'- and 3'-exon (927 nt) [22], the 5'-fragment of HCV genome (strain Jc1, residues 1–1271)[23], HOTAIR D1 (532 nt) [24] and *Oceanobacillus iheyensis* Group II intron with 5'- and 3'-exon (803 nt) [25]. The primer sequences are TAATAGGTGAGGTTTCAATG for RepA D3, TATCAGTTATATATAACC for Group II intron ai5γ, TTGCAGTCTTGGACAAACC for HCV (1–1236), AATTCTCTCGCCAATGTGC for HOTAIR D1 and TACTATGACTTCTGCTGAC for *O. iheyensis* Group II intron respectively. Corresponding primers designed to carry RT reactions were annealed to positions 1630 of RepA D3, position 764 of Group II intron ai5γ, position 1254 of HCV, position 434 of HOTAIR D1 and position 573 of *O. iheyensis* Group II intron respectively.

Each RT primer was labeled with 32P at the 5’-end by T4 PNK and purified on a 12% polyacrylamide gel as described previously[2]. The final concentration was 100 nM for each RNA template and 500 nM for MarathonRT. The initial reaction buffer for MarathonRT contained 50 mM K-HEPES pH 8.5, 100 mM KCl, 2 mM MgCl₂ and 5 mM DTT, and the optimal reaction buffer contained 50 mM Tris-HCl pH 8.3, 200 mM KCl, 2 mM MgCl₂, 5
mM DTT and 20% glycerol. The other buffer conditions are specifically described in the main text. Reactions were carried out at 42°C for 15 minutes. The RT reactions were stopped by incubation at 95°C for 1 min. After cooling down, the RT enzymes were digested by protease K at 42°C for 15 min, and RNA templates were hydrolyzed by 300 mM NaOH at 95°C for 5 min.

To analyze radiolabeled cDNA products, they were first resolved on a 5% polyacrylamide denaturing gel that contained 7 M urea; the gel was then dried at 80°C for 30 min under vacuum; finally, the dried gel was exposed to a phosphorimager screen that was subsequently scanned by Typhoon FLA 9500 scanner (GE Healthcare). The resulting image was analyzed and quantified by ImageQuant software (GE Healthcare).

**Thermal shift assay**

Thermal shift assay reactions (20 μl) contained 1x GloMelt dye from Biotium Inc, 0.3 mg/mL purified MarathonRT protein and 1x each of the reaction buffers, as described in the main text. The reactions were carried out in a 96-well plate using a Bio-Rad C1000 real-time thermocycler to monitor the fluorescence change in the FAM channel, according to the protocol from Biotium Inc. To scan a wide range of temperature, all the reactions were initially set up on ice. To initiate the thermal shift assay, the reactions were held at 10°C for 2 min and then the temperature was increased to 80°C at the ramp rate of 0.5°C/15 sec. Each of the reactions was repeated 4 times and averaged to generate the protein melting curve plot.

**HCV RNA transcription, folding, and modification**

Template plasmid was linearized with XbaI and purified using a Qiagen PCR Cleanup kit. The HCV RNA genome (Jc1)[26] was *in vitro* transcribed by run-off transcription using T7 RNA Polymerase P266L variant[27]. Following transcription, DNA was degraded with RQ1 DNase (Promega) followed by addition of Proteinase K (ThermoFisher) to degrade enzymes. RNA was treated with 25 mM final EDTA pH 8.0 to chelate Mg^{2+} and then concentrated using 100 kDa Amicon Ultra filtration columns at room-temperature. Size-exclusion chromatography was performed at room-temperature using a self-packed Sephacryl S-1000 column with a 24 mL bed volume, equilibrated with Filtration Buffer (50 mM K-HEPES pH 7.5, 150 mM KCl, 100 μM EDTA pH 8.0). RNA from the peak fraction was diluted to 250 ng/μL, MgCl₂ was added to a final concentration of 10 mM, and RNA was folded at 37°C for 30 minutes. RNA probes were added to a final concentration of 10 mM 1M7, 200 mM NAI, 300 mM FAI, and 0.4% DMS, and incubated for 10 minutes at 37°C. RNA was then LiCl precipitated, washed with 80% EtOH, and resuspended in ME Buffer (10 mM MOPS pH 6.0 and 1 mM EDTA).

**Reverse Transcription and Library Preparation for MaP**

*In vitro* transcribed RNA (1 μg) was reverse transcribed using 40 U MarathonRT in optimized RT Buffer (50 mM Tris-HCl pH 8.3, 200 mM KCl, 5 mM DTT, 20% glycerol, 1 mM MgCl₂ or 1 mM MnCl₂) and 0.5 mM dNTP mix using primer TTTTTTCTTTGAGGTAGTAGG. For SSII MaP, RNA was reverse transcribed using 200 U SSII using the recommended buffer conditions (50 mM Tris-HCl pH 8.3, 75 mM KCl, 10
mM DTT, 6 mM MnCl$_2$, 0.5 mM dNTP mix$[^3]$. RT reactions were incubated at 42°C for 3 hours. Following reverse transcription, EDTA was added to 25 mM final and cDNA was precipitated with 300 mM sodium acetate pH 5.2 and 80% EtOH. An amplicon covering the HCV 5' UTR was generated from cDNA using Q5 Hot Start Polymerase (NEB) and the following primers: CTAATAGGGGCGACACTCCG, CTTTGAGGTAGATTTGTGCTC. Amplicons were purified with Monarch DNA Cleanup Kits (NEB) and diluted to 0.2 ng/μL. Sequencing libraries were generated using a Nextera XT DNA Library Preparation Kit (Illumina) according to manufacturer instructions, but with 1/5th the recommended volume. Libraries were quantified using a Quibit (Life Technologies) and Bioanalyzer (Agilent) and sequenced on a NextSeq 500/550 (Illumina).

**Harvesting, reverse transcription, and library prep from cellular RNA**

For *in vivo* SHAPE-MaP, Huh7.5 cells constitutively expressing the HCV subgenomic replicon (SGR-neo-JFH-1) were grown to 80% confluence. The cells were washed with PBS, collected by centrifugation at 1,500 rpm for 5 minutes and resuspended in PBS. A final concentration of 200 mM NAI or the same amount of DMSO was added to the cells. The reaction was incubated at room temperature for 10 minutes. Then, RNA was extracted by the addition of Trizol according to manufacturer recommendations and then purified with Qiagen RNeasy kit. Then, RNA was resuspended in ME buffer to a final concentration of 1 μg/μL. RNA (1 μg) was reverse transcribed by MarathonRT using the optimized reaction buffer for *in vivo* RNA (50 mM Tris-HCl pH 7.5, 200 mM KCl, 5 mM DTT, 20% glycerol, 1 mM MnCl$_2$ and 0.5 mM dNTP mix) or by SSII as described in the text. The reverse transcription reactions were carried out at 42°C for 3 hours using primer TTTTTTCTTTGAGGTAGTAGG for short amplicon and primer TAATGTTGGGATTGATGCC for long amplicon. Following reverse transcription, the cDNA products were cleaned up using Agencourt AMPure XP beads, and amplified by Q5 Hot Start Polymerase (NEB) with primers CTAATAGGGGCGACACTCCG and CTTTGAGGTAGATTTGTGCTC for short amplicon and primers CTAATAGGGGCGACACTCCG and GTTGGATTGATGCCATGTG for long amplicon. The following library preparation steps were the same as described above.

For natural modification detection, Huh7.5 cells constitutively expressing the same HCV subgenomic replicon were grown to 80% confluence. RNA was extracted by adding Trizol followed by washing cells with PBS according to manufacturer recommendations. RNA was resuspended in ME buffer to a final concentration of 1 μg/μL. RNA (1 μg) was reverse transcribed by MarathonRT as above using the optimized RT buffer with indicated MgCl$_2$ or MnCl$_2$ concentrations or by SSII as above. For HCV 5' UTR analysis, the protocol above was followed. For 18S and 28S analysis, reverse transcription was performed using random hexamers (LifeTech) followed by second strand synthesis using NEBNext Ultra II Non-Directional Second Strand Synthesis Module (NEB). Double-stranded DNA was purified with Monarch DNA Cleanup Kits (NEB) and diluted to 0.2 ng/μL. Libraries were generated and sequenced as above.
Computational analysis of MaP and mutational Data

All libraries were analyzed using ShapeMapper 2[14], aligning to the HCV JFH1 (accession number: AB047639.1) 5' UTR, nucleotides 1–397, 18S or 28S RNA using default settings. HCV IRES structure prediction was performed using nucleotides 43–353 and corresponding reactivities with ShapeKnots[28]. Sensitivity and Positive Predictive Value (PPV) were calculated by comparing predictions output by ShapeKnots with the published crystal structure of the HCV IRES and equations 1 and 2, respectively.

\[
\text{Sensitivity} = \frac{\text{number of correctly predicted base pairs}}{\text{total number of known base pairs}} \times 100
\]

\[
\text{PPV} = \frac{\text{number of correctly predicted base pairs}}{\text{total number of predicted base pairs}} \times 100
\]

HCV 5' UTR structures and reactivities were drawn using RNAstructure[29]. The significance of mutation rates induced by natural modifications was analyzed using equal variance t-test. Before performing t-test, the mutation rates were log_{10}-transformed to abide by normality.

Analysis of Mutation Type Bias

For both treated and untreated samples, mutation counts were output by passing the “–output-counted-mutations” flag when calling the ShapeMapper2 script. For every nucleotide, raw mutation counts were normalized to the effective read depth at that nucleotide (Eq. 3) to determine the probability of a specific mutation type (i.e., insertions, deletions, etc.).

\[
P_{raw}(\text{mut}) = \frac{n_{\text{mutation type},i}}{\text{depth}_i}
\]

For SHAPE applications, mutation probabilities derived from untreated/unmodified samples were subtracted from the mutation probabilities determined for modified samples (Eq. 4). This ensured that, when making comparisons between enzymes, only mutations arising from acylated nucleotides were considered in downstream analysis steps. No subtraction was performed for analysis of mutations induced by natural modifications.

\[
P(\text{mut}) = P_{\text{treated},i}(\text{mut}) - P_{\text{untreated},i}(\text{mut})
\]

Data was normalized to determine distributional bias of specific mutation types, given that a mutation has been made. To determine the mutational bias arising as a function of the base modification and the RT enzyme used, mutations were binned by type (insertions, deletions, substitutions, multinucleotide mismatches, and complex mutations), and normalized to the probability that any mutation is made (Eq. 5)
\[ \frac{P_{type}(\text{mut})}{\sum P_{all}(\text{mut})} \]  

Eq. 5

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Chen Zhao (Rockefeller University, prior Pyle lab member) for her suggestions on MarathonRT purification and reaction optimization.

Funding

This work was supported by NIH 1 R01 HG009622.

Abbreviations

- RT: Reverse transcriptase
- IncRNA: long noncoding RNA
- SHAPE-MaP: selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling
- cDNA: complementary DNA
- DMS-MaP: dimethyl sulfate mutational profiling with sequencing
- SSII: SuperScript II
- MMLV: Molony Murine Leukemia Virus
- BSA: bovine serum albumin
- TGIRT: thermostable group II reverse transcriptase
- HCV: hepatitis C virus
- IRES: internal ribosome entry site
- RING-MaP: RNA interaction groups by mutational profiling

REFERENCES

[1]. Zhao C, Pyle AM. Crystal structures of a group II intron maturase reveal a missing link in spliceosome evolution. Nat. Struct. Mol. Biol. 23 (2016) 558–565. [PubMed: 27136328]
[2]. Zhao C, Liu F, Pyle AM. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. RNA 24 (2018) 183–195. [PubMed: 29109157]
[3]. Siegfried NA, Busan S, Rice GM, Nelson JA, Weeks KM. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat. Methods 11 (2014) 959–965. [PubMed: 25028896]
[4]. Zubradt M, Gupta P, Lambowitz AM, Weissman JS, Rouskin S. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. Nat. Methods 14 (2017) 75–82. [PubMed: 27819661]

[5]. Gerard GF, Potter RJ, Smith MD, Rosenthal K, Dharival G, Lee J, et al. The role of template-primer in protection of reverse transcriptase from thermal inactivation. Nucleic Acids Res. 30 (2002) 3118–3129. [PubMed: 12136094]

[6]. Spiess AN, Ivell R. A highly efficient method for long-chain cDNA synthesis using trehalose and betaine. Anal. Biochem. 301 (2002) 168–174. [PubMed: 11814287]

[7]. Schrader C, Schielke A, Ellerbroek L, Johne R. PCR inhibitors - occurrence, properties and removal. J. Appl. Microbiol. 113 (2012) 1014–1026. [PubMed: 22779764]

[8]. Strutzenberg TS, Garcia-Ordonez RD, Novick SJ, Park H, Chang MR, Doeblin C, et al. HDX-MS reveals structural determinants for RORgamma hyperactivation by synthetic agonists. Elife 8 (2019) e47172.

[9]. Sexton AN, Wang PY, Rutenberg-Schoenberg M, Simon MD. Interpreting Reverse Transcriptase Termination and Mutation Events for Greater Insight into the Chemical Probing of RNA. Biochemistry 56 (2017) 4713–4721. [PubMed: 28820243]

[10]. Adams RL, Huston NC, Tavares RCA, Pyle AM. Sensitive detection of structural features and rearrangements in long, structured RNA molecules. Methods Enzymol. 623 (2019) 249–289. [PubMed: 31239050]

[11]. Beckman RA, Mildvan AS, Loeb LA. On the fidelity of DNA replication: manganese mutagenesis in vitro. Biochemistry 24 (1985) 5810–5817. [PubMed: 3910084]

[12]. Quade N, Boehringer D, Leibundgut M, van den Heuvel J, Ban N. Cryo-EM structure of Hepatitis C virus IRES bound to the human ribosome at 3.9-A resolution. Nat. Commun. 6 (2015) 7646. [PubMed: 26155016]

[13]. Smola MJ, Rice GM, Busan S, Siegfried NA, Weeks KM. Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. Nat. Protoc. 10 (2015) 1643–1669. [PubMed: 26426499]

[14]. Busan S, Weeks KM. Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. RNA 24 (2018) 143–148. [PubMed: 29114018]

[15]. Sutlovic D, Definis Gojanovic M, Andelinovic S, Gugic D, Primorac D. Taq polymerase reverses inhibition of quantitative real time polymerase chain reaction by humic acid. Croat. Med. J. 46 (2005) 556–562. [PubMed: 16100758]

[16]. Mohr S, Ghanem E, Smith W, Sheeter D, Qin Y, King O, et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. RNA 19 (2013) 958–970. [PubMed: 23697550]

[17]. Homan PJ, Favorov OV, Lavender CA, Kursun O, Ge X, Busan S, et al. Single-molecule correlated chemical probing of RNA. Proc. Natl. Acad. Sci. USA 111 (2014) 13858–13863. [PubMed: 25205807]

[18]. Lempereur L, Nicoloso M, Riehl N, Ehresmann C, Ehresmann B, Bachellerie JP. Conformation of yeast 18S rRNA. Direct chemical probing of the 5′ domain in ribosomal subunits and in deproteinized RNA by reverse transcriptase mapping of dimethyl sulfate-accessible. Nucleic Acids Res. 13 (1985) 8339–8357. [PubMed: 24171979]

[19]. Helm M, Motorin Y. Detecting RNA modifications in the epitranscriptome: predict and validate. Nat. Rev. Genet. 18 (2017) 275–291. [PubMed: 28216634]

[20]. Tserovski L, Marchand V, Hauenschild R, Blanloeil-Oillo F, Helm M, Motorin Y. High-throughput sequencing for 1-methyladenosine (m(1)A) mapping in RNA. Methods 107 (2016) 110–121. [PubMed: 26922842]

[21]. Liu F, Somarowthu S, Pyle AM. Visualizing the secondary and tertiary architectural domains of lncRNA RepA. Nat. Chem. Biol. 13 (2017) 282–289. [PubMed: 28068310]

[22]. Fedorova O, Solem A, Pyle AM. Protein-facilitated folding of group II intron ribozymes. J. Mol. Biol. 397 (2010) 799–813. [PubMed: 20138894]

[23]. Pirakitikul N, Kohlway A, Lindenbach BD, Pyle AM. The Coding Region of the HCV Genome Contains a Network of Regulatory RNA Structures. Mol. Cell 62 (2016) 111–120. [PubMed: 26924328]
[24]. Somarowthu S, Legiewicz M, Chillon I, Marcia M, Liu F, Pyle AM. HOTAIR forms an intricate and modular secondary structure. Mol. Cell 58 (2015) 353–361. [PubMed: 25866246]

[25]. Toor N, Keating KS, Taylor SD, Pyle AM. Crystal structure of a self-spliced group II intron. Science 320 (2008) 77–82. [PubMed: 18388288]

[26]. Pietschmann T, Kaul A, Koutsoudakis G, Shavinskaya A, Kallis S, Steinmann E, et al. Construction and characterization of infectious intragenotypic and intergenotypic hepatitis C virus chimeras. Proc. Natl. Acad. Sci. USA 103 (2006) 7408–7413. [PubMed: 16651538]

[27]. Tang GQ, Nandakumar D, Bandwar RP, Lee KS, Roy R, Ha T, et al. Relaxed rotational and scrunching changes in P266L mutant of T7 RNA polymerase reduce short abortive RNAs while delaying transition into elongation. PLoS One 9 (2014) e91859.

[28]. Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. Proc. Natl. Acad. Sci. USA 110 (2013) 5498–5503. [PubMed: 23503844]

[29]. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 11 (2010) 129. [PubMed: 20230624]
Figure 1.  
Incrementally optimizing the buffer components for MarathonRT. (A) In the first step, primer extension efficiency was measured in Tris-HCl buffer to identify the pH optimum. (B) Then the effect of MgCl\(_2\) concentration was titrated to determine the optimum. (C) Subsequently, the effect of KCl concentration was examined in the range of 50 mM to 300 mM KCl. (D) Finally, a few protein stabilizing agents were further tested in the incrementally optimized buffer that contains 50 mM Tris-HCl pH 8.3, 200 mM KCl, 2 mM MgCl\(_2\) and 5 mM DTT. The protein stabilizing agents include 1 M betaine (bet.), 0.6 M...
trehalose (tre.) and 20% glycerol (gly.). (E) Primer extension efficiency was measured in the suboptimal buffer (no add.) containing 50 mM Tris-HCl pH 7.5, 200 mM NaCl, 2 mM MgCl$_2$ and 5 mM DTT. Primer extension with addition of 1 M betaine (bet.), 0.6 M trehalose (tre.) or 20% glycerol (gly.) was measured separately. (F) The gel figure for representative conditions in (A) – (E) labeled with numbers. The reaction condition denoted by a red arrow in (A), (B) and (C) was determined as optimal and forwarded to the next round of optimization. All experiments were repeated three times with means and standard deviations shown in the figures.
Figure 2.
Exploring the thermal stability of MarathonRT in representative buffers. (A) Primer extension efficiency was measured to examine protein stability at 47°C and 50°C in the representative buffers, including initial buffer (ini.), the optimized buffer without protein stabilizing agent (no add.), with 1 M betaine (bet.), 0.6 M trehalose (tre.), 1 M betaine and 0.6 M trehalose (bet.+tre.) and 20% glycerol (gly.). Each reaction was carried out in triplicate with mean and standard deviation shown in the figures. (B) Thermal shift assays were performed in the same buffers. The curves were generated from the averaged result of four repeats for each buffer.
Figure 3.
Marathon RT incorporates mutations at probed nucleotides on an HCV RNA template. (A) MarathonRT efficiently generated full-length cDNA at 0.5 – 2 mM MnCl$_2$ using the DMS modified full-length HCV template. The experiment was performed twice. (B) MarathonRT incorporated more mutations on 1M7, NAI, FAI, and DMS modified HCV IRES templates in 1 mM MnCl$_2$. The increased mutation rates mostly resulted from single-stranded nucleotides. The experiment was performed three times to confirm the reproducibility.
Figure 4.
The comparison of mutational type distribution by MarathonRT with 1 mM Mn$^{2+}$ and SSII with 6 mM Mn$^{2+}$ in DMS-/SHAPE-MaP on modified HCV IRES RNA templates. The overall mutational type distribution with SHAPE reagents 1M7, DMS, NAI and FAI for (A) MarathonRT and (B) SSII. DMS-MaP mutation type distributions are further detailed for (C) MarathonRT and (D) SSII. A, C, U, and G slices refer to the frequency at which substitutions occurred at those nucleotides. “Other” includes all other mutation types (deletions, insertions, multinucleotide mismatches, and complex mutations).
Figure 5.
In-cell SHAPE-MaP reactivity with NAI can be used to accurately predict RNA secondary structure from short and long amplicons (2.5 kb) generated by MarathonRT. (A) A sketch showing where amplicons were generated on the HCV subgenome. (B) MarathonRT-MaP resulted in higher mutation rates at modified nucleotides in HCV-IRES single stranded regions by NAI when using a reaction buffer with 1mM Mn$^{2+}$ and a pH of 7.5. * $P<0.05$ (unpaired two-tailed Student’s t-test, performed using log$_{10}$ transformed mutation rate). (C) NAI reactivity, calculated from the mutation rate of MarathonRT during reverse transcription on the modified RNA template with reaction conditions described above, mapped to the HCV IRES structure. This experiment was repeated three times to confirm reproducibility.
Figure 6.
MarathonRT is sensitive to certain natural modifications on rRNA with Mn\(^{2+}\) as the cofactor. (A) In the presence of Mn\(^{2+}\), 2’-OMe induced dramatically increased mutation rate compared to the unmodified nucleotides. *** \(P<0.001\) (unpaired two-tailed Student’s t test, performed using log scale of mutation rate). (B) Pseudouridine induced mildly increased mutation rate at higher Mn\(^{2+}\) concentrations. ns=not significant, * \(P<0.05\), ** \(P<0.01\) (unpaired two-tailed Student’s t test, performed using log scale of mutation rate). (C) Plotting of mutation rates for each nucleotide on 18S rRNA and (D) 28S rRNA collected with 2mM Mn\(^{2+}\) or 1mM Mg\(^{2+}\) as the RT cofactor. The experiment was repeated three times to confirm reproducibility.