Long-read transcriptome-wide RNA structure maps using DMS-FIRST-seq

Oguzhan Begik[1], Gregor Diensthuber[1,2], Ivana Borovska[3], John S Mattick[4], Danny Incarnato[3] and Eva Maria Novoa[1,2,5,#]

[1]Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain

[2] Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain.

[3] Department of Molecular Genetics, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, Groningen, the Netherlands.

[4] School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, Sydney NSW 2052, Australia.

[5] ICREA, Pg. Lluis Companys 23, Barcelona 08010, Spain

[#] Correspondence to: Eva Maria Novoa (eva.novoa@crg.eu)

ABSTRACT

Coupling chemical probing with nanopore direct RNA sequencing (dRNA-seq) has been proposed as a promising approach to capture RNA structural dynamics at single-molecule resolution. However, dRNA-seq produces widespread rather than single-nucleotide signals from modified bases, thereby limiting its precision in RNA structure prediction. Here we introduce FIRST-seq, a method that overcomes these limitations by combining chemical probing with nanopore cDNA sequencing to achieve single nucleotide resolution. By systematically assessing various RT enzymes and buffers, we optimize conditions that minimize drop-off and enhance error signatures linked to RNA structure. Our results demonstrate that DMS-FIRST-seq generates high-resolution, accurate RNA structure maps in both *in vivo* and *in vitro* contexts, thus providing a simple and cost-effective method for transcriptome-wide isoform-specific RNA structural analyses.

INTRODUCTION

In recent decades, it has become clear that RNAs are not simple intermediary molecules between DNA and proteins, but are in fact functional molecules playing key roles in regulating cellular and developmental processes, including genome organization, gene expression, and splicing [1–4]. The single-stranded nature and 2' hydroxyl group of RNA promote intramolecular base-pairing and structure formation, which in turn governs its function. Therefore, accurate RNA structural maps are crucial for understanding the complexity, function, and regulation of RNA molecules[5].

Dimethyl sulfate (DMS) and 2'-hydroxyl acylation and primer extension (SHAPE) reagents have traditionally been used to obtain experimental measurements on base-pairing and tertiary interactions of the RNA molecules[6–8]. Coupling of DMS and SHAPE chemical labeling with next-generation sequencing has provided genome-wide RNA structure maps that can inform on the dynamics of RNA structures in a variety of cellular contexts[9–12]. While these approaches have provided high accuracy information of RNA structures in a transcriptome-wide fashion, they also have several caveats: (i) the short-read nature of these methods limits their resolution, which cannot provide per-isoform RNA structure maps; (ii) they are unable to produce maps for highly repetitive regions, as short reads are unmappable, and (iii) they do not provide information of single RNA molecules, but rather of an ensemble of them.

To overcome these limitations, several groups have proposed the use of chemical probing coupled to third-generation sequencing technologies, such as PacBio[13] or Oxford Nanopore Technologies (ONT)[14–16]. While both PacBio and ONT are capable of producing full-length reads from cDNA molecules, only the latter allows sequencing of native RNA molecules (dRNA-seq)[14–16], as it can detect both native[17–19] and chemical-induced RNA modifications[15], captured in the form of disruptions in the measured ionic current as the molecule passes through the pore[17–19].

A common chemical probing reagent is dimethyl sulfate (DMS), which methylates unpaired adenosines (A) and cytosines (C), resulting in chemically modified nucleosides *N1*-methyladenosine (m^1A), *N3*-methylcytosine (m^3C) respectively (**Figure 1a**). However, DMS also methylates both unpaired and paired guanosines (G) to form N7-methylguanosine (m^7G) modifications, which can confound the prediction of RNA structures, due to the 'leak' of modification information to neighboring positions[19,20], resulting in false positive predictions with higher accessibility (**Figure 1b**). Moreover, G positions will be methylated regardless of their pairing status, leading to higher background errors.

As an alternative, recent studies have shown that coupling DMS probing with nanopore cDNA sequencing (dcDNA-seq) yields improved RNA structure prediction maps of viral transcripts, when used with highly processive group II intron reverse transcriptases such as Marathon RT (Nano-DMS-MaPseq)[16]. This method however, relies on the use of gene-specific primers, and thus cannot be applied transcriptome-wide.

Here we provide a novel method to obtain transcriptome-wide RNA structure maps, by using DMS probing coupled to nanopore cDNA sequencing. The method relies on FIRst STrand sequencing (FIRST-seq), which does not require second-strand synthesis nor PCR amplification, thus simplifying the cost and time of library preparation while allowing for transcriptome-wide capture of RNA transcripts, overcoming the need of gene-specific primers[16]. We systematically examine the performance of a battery of reverse transcriptase enzymes, under different buffer conditions, to identify the combination that yields highest reverse transcription mismatch error signatures at m^1A and m^3C sites, while minimizing the reverse transcription drop-off. We then apply FIRST-Seq to both *in vitro* transcribed DMS-probed constructs as well as *in vivo* DMS-treated cells, demonstrating that FIRST-seq recapitulates known RNA structures with high accuracy.

RESULTS

Direct RNA sequencing data provides more wide-spread signal of the modified nucleotide

By combining dRNA-seq with chemical probing techniques, it has become possible to investigate RNA structures with isoform-specific resolution[14]. While previous works have been successful at capturing RNA structure information, we reasoned that dRNA-seq might not be optimal at capturing chemical-probing signals, as these are known to cause disruptions in current intensity and basecalling 'errors' not only at the modified site of interest, but also at neighboring positions, leading to higher background error and decreased resolution[15,18,19,21].

To overcome this problem, we reasoned that cDNA-based approaches coupled to long read sequencing might offer a plausible solution to capture RNA structure information at single nucleotide, single molecule resolution, despite requiring a reverse-transcription step. To examine whether cDNA-based signatures of chemical-induced modifications would show higher resolution than dRNA-seq-based signatures, we first analyzed the RNA modification patterns using yeast rRNA and tRNA molecules, for which a vast amount of RNA modification sites are well annotated (**Figure 1c,** see also **Figure S1a**), and sequenced them using both dRNA-seq and dcDNA-seq. Our results show that WC-disrupting RNA modifications, including m^1A and m^3C, show single nucleotide resolution information in dcDNA-seq datasets, whereas WC-neutral RNA modified sites do not cause detectable variations in the base-calling patterns (**Figure 1d,** see also **Figure S1b**). Moreover, the RT misincorporation 'error' signatures observed in dcDNA-seq were significantly stronger than those observed in the form of 'basecalling errors' in dRNA-seq data at the modified site in question, which can be used as a proxy for signal intensity differences (**Figure 1e,** see also **Figure S1b,c**), suggesting that cDNA-based approaches offer a more robust approach to identify chemical-induced modifications, such as m^3C and m^1A, compared to dRNA-seq (**Table S1** and **S2**).

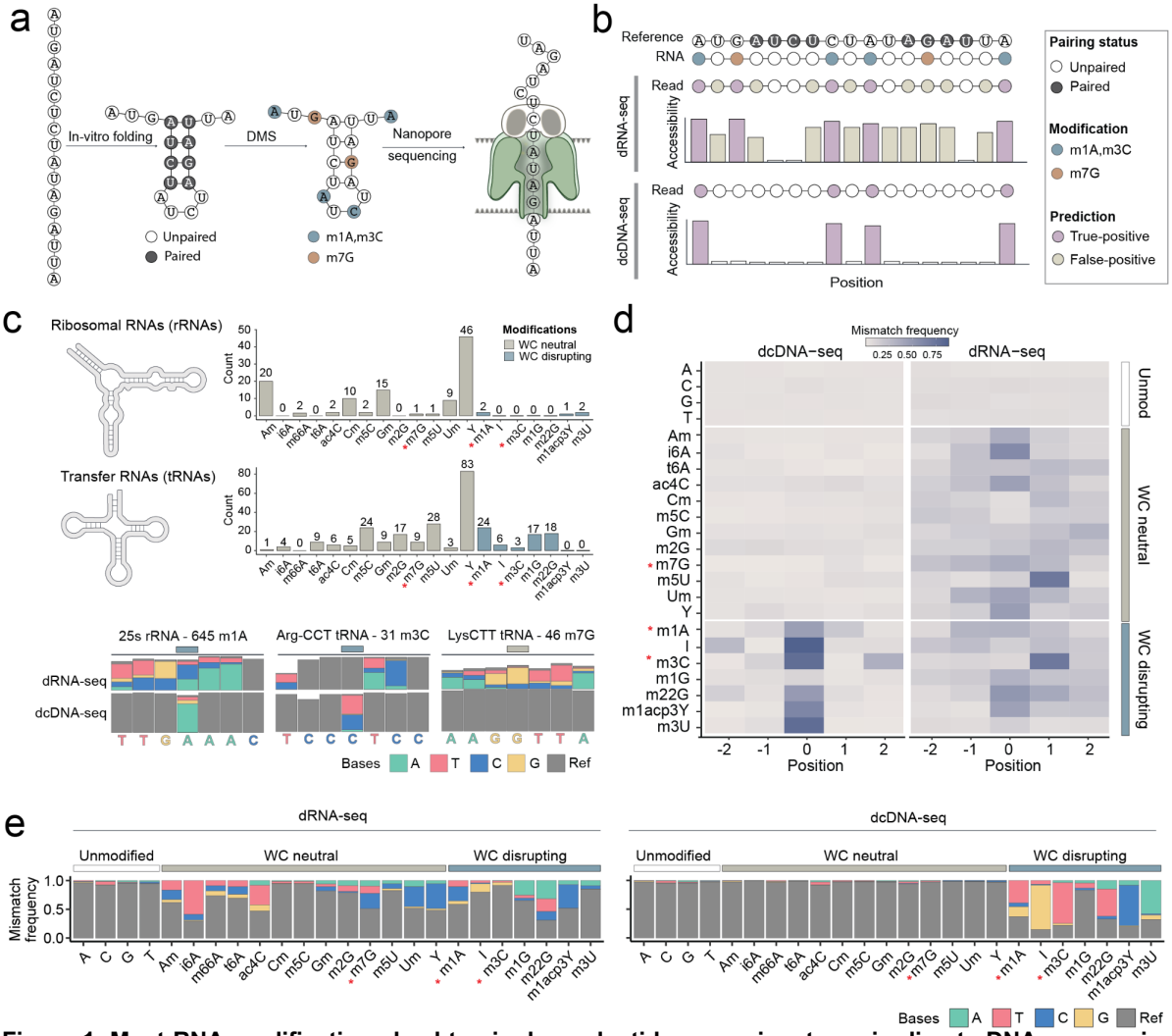


Figure 1. Most RNA modifications lead to single nucleotide error signatures in direct cDNA sequencing, whereas they produce 'spread' error signals in direct RNA sequencing datasets. (a) Schematic overview of the DMS probing coupled to nanopore sequencing, which can be sequenced either in its native RNA form (dRNA-seq) or after reverse transcription (dcDNA-seq). Gray-colored circles indicate paired bases, while white circles represent unpaired bases. Dimethyl sulfate (DMS) treatment methylates unpaired A and C positions, as well as both paired and unpaired G positions. Methylation leads to the formation of m^1A and m^3C modifications (blue) and m^7G modifications (orange). **(b)** Schematic representation of how DMS-induced modifications affect the signal in the RNA molecule and consequently the mapped accessibility. While dRNA-seq detects modified sites (shown in purple), the "error" signature often spans to neighboring unmodified bases (shown in brown), resulting in a higher false positive rate. In contrast, direct cDNA sequencing (dcDNA-seq) produces "error" signatures only from WC-disrupting modifications (m^1A, m^3C), ignoring m^7G, which is modified regardless of its pairing status. Consequently, cDNA reads from DMS-probed RNAs are expected to have a higher true positive error rate. **(c)** Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) are heavily modified (top panel). In the bottom panel, Integrated Genomics Viewer (IGV) coverage track snapshots exemplify dRNA-seq and dcDNA-seq data at three modified positions Colored bars on top of the barplots indicate the modified positions. See also **Figure S1a** for additional IGV snapshots. (**d)** Heatmap showing aggregated basecalling error signatures along the 5-mer region (centered at the modified site, position 0) from multiple modified positions in tRNAs and rRNA molecules, sequenced with either dRNA-seq or dcDNA-seq. **See also Figure S1b,c**. **(e)** Barplot showing aggregated error signatures using dRNA-seq (left panel) and dcDNA-seq (right panel) data. DMS-induced modifications (m^1A, m^3C and m^7G) are marked with a red star.

FIRST-seq expands nanopore cDNA sequencing compatibility to all reverse transcriptases

The commercial direct cDNA sequencing (dcDNA-seq) and PCR-cDNA sequencing protocols offered by ONT ensure capturing full-length cDNA products by requiring a "strand-switching" step, followed by second-strand synthesis[22]. Double-stranded cDNA products (or its PCR-amplified equivalents) can then be ligated to ONT adapters (**Figure 2a**, left panel). While these methods have proven useful for RNA quantification and isoform detection[23,24], their applicability for RNA structure determination is limited, as they rely on the use of reverse transcriptases (RT) such as Maxima H Minus RT (Maxima) that adds untemplated GGG overhangs at the 3' end of the cDNA strands, which in turn is used to initiate the strand-switching and second-strand synthesis. Consequently, this limits the possibility of using highly processive RT enzymes, such as group II intron RT enzymes – which are reported to outperform viral RTs for RNA structure prediction[9,16] due to their improved processivity and increased misincorporation rates[25] - in cDNA-based protocols coupled to nanopore sequencing. This limitation was recently overcome by using primer-specific oligonucleotides coupled to PCR amplification, which allowed the authors to use Marathon RT enzyme to generate Nano-DMS-MaP-seq libraries[9,16]. However, a major caveat of this method is that it will be limited to individual transcripts that are complementary to the gene-specific oligonucleotides used to initiate the reverse transcription step, and thus cannot be applied transcriptome-wide.

To address this gap, here we developed FIRST-seq (**Figure 2a**, right panel), a method that allows sequencing the first strand cDNA products –including truncated RT products– transcriptome-wide, without the need for second strand synthesis nor PCR amplification. Notably, FIRST-seq can be coupled to any reverse transcriptase of interest, as it does not require RT enzymes to leave a specific overhang at the 3' end of the cDNA to perform second strand synthesis. Moreover, it captures reverse-transcription drop-offs (RT drop-off) at positions that coincide with RNA modifications (**Figure S1d,e**), in addition to capturing information arising from basecalling errors (misincorporations, deletions and insertions) (**Figure S1f,g**).

Figure 2 (*legend in next page*)

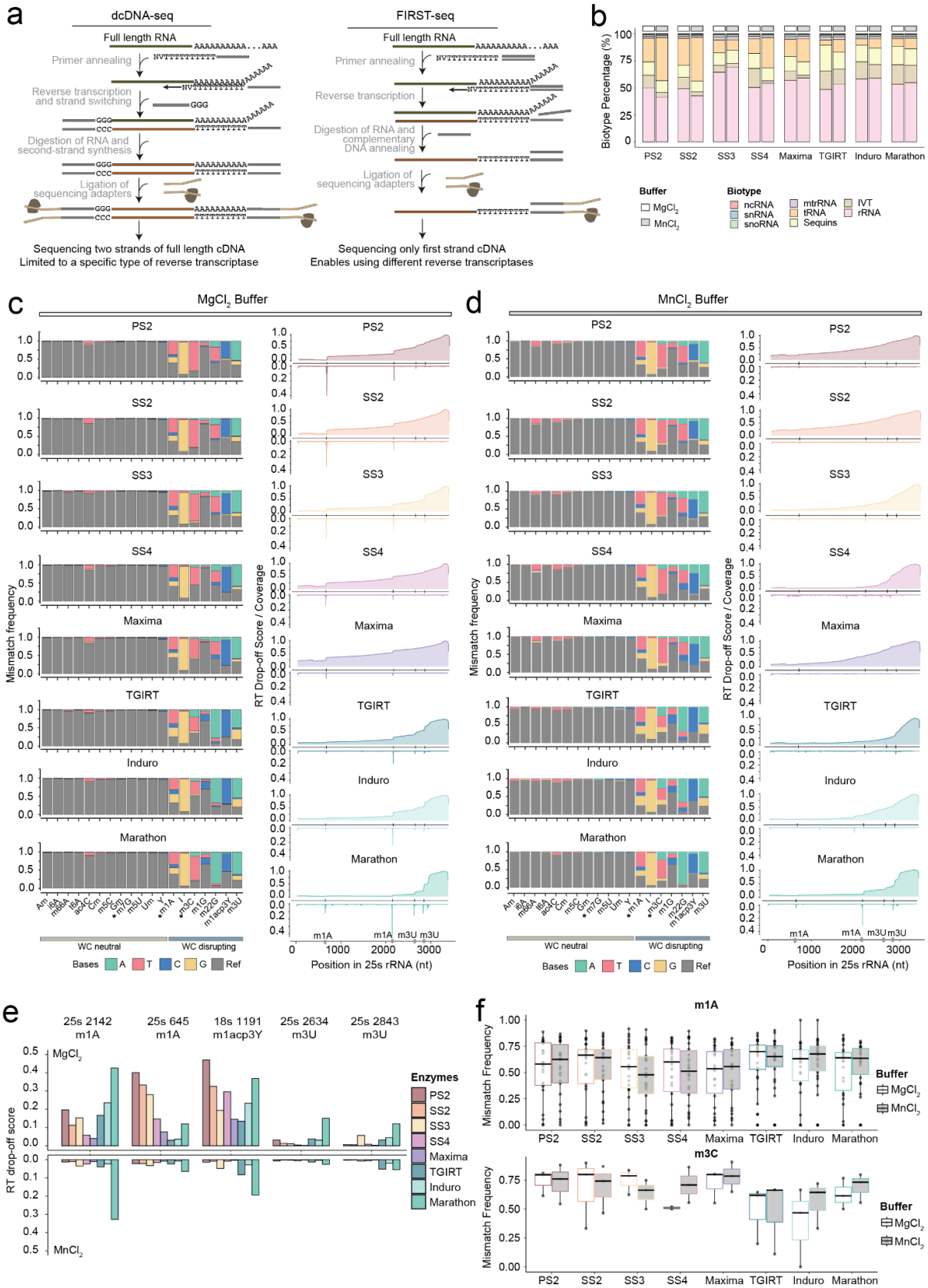


Figure 2. Comparison of reverse transcriptase (RT) performance in $MgCl_2$ and $MnCl_2$ buffers across different enzymes. (a) Schematic comparing workflows of dcDNA-seq and FIRST-seq. FIRST-seq simplifies the process by sequencing only the first cDNA strand, eliminating the need for strand-switching and second-strand synthesis. This allows for the use of a broader range of reverse transcriptases and avoids strand-switching artifacts. dcDNA-seq, on the other hand, captures both strands, but is limited by the requirement of strand-switching for full-length cDNA synthesis. **(b)** Stacked barplot showing the biotype percentage of reads generated by each RT enzyme in either magnesium ($MgCl_2$) or manganese ($MnCl_2$) buffer. Biotypes include non-coding RNAs (ncRNA), ribosomal RNAs (rRNA), small nucleolar RNAs (snoRNA), mitochondrial RNAs (mtRNA), transfer RNAs (tRNA), spike-in RNAs (Sequins), and structured RNA IVTs (IVT) **(c,d)** Barplots (left panels) showing mismatch frequency across Watson-Crick (WC)-neutral and WC-disrupting bases for each RT enzyme in $MgCl_2$ (d) and $MnCl_2$ (e) buffers. The bar colors represent different base types: adenine (A), thymine (T), cytosine (C), guanine (G), and reference bases (gray). Line plots (right panels) depict RT drop-off scores as a function of position in the 25S rRNA, with WC-disrupting modifications (m^1A, m^3U, m^3C) highlighted. **(e)** Barplot showing RT drop-off scores at specific 25S rRNA modification sites (m1A at positions 2142 and 645, m^1acp^3Y at position 1191, m^3U at positions 2634 and 2843) for each RT enzyme in $MgCl_2$ (top) and $MnCl_2$ (bottom) buffers. Colors correspond to the RT enzymes as indicated in the legend. **(f)** Boxplots showing mismatch frequency distributions at m^1A (top) and m^3C (bottom) modification sites across the RT enzymes in $MgCl_2$ and $MnCl_2$ buffers. Buffer conditions are represented by color coding, with $MgCl_2$ in light gray and $MnCl_2$ in dark gray. For the boxplots, the box extends from the first quartile to the third quartile of the data, with a line at the median. The whiskers extend from the box to the farthest data point lying within 1.5x the interquartile range from the box.

Systematic benchmarking of RT drop-off and mismatch signatures across RT enzymes and buffer conditions

Compared to other cDNA-based methods, FIRST-seq allows the use of distinct reverse transcriptases by bypassing the strand-switching step (**Figure 2a**). Here, we exploited this feature to examine which RT enzymes and buffer conditions might enhance the detection of chemical-induced and natural RNA modifications in the form of systematic basecalling errors and/or RT drop-offs using FIRST-seq. To this end, we systematically evaluated the performance of 8 different RT enzymes (Protoscript, Superscript II (SS2), Superscript III (SS3), Superscript IV (SS4), Maxima RT H Minus (Maxima), TGIRT, Induro RT (Induro) and Marathon RT (Marathon)) (**Table S3**) on a mix of RNAs (*S. cerevisiae* total RNA and several *in vitro*-transcribed RNAs: *B. subtilis* guanine riboswitch, *B. subtilis* lysine riboswitch, *Tetrahymena* ribozyme and sequin RNAs) **(Figure S2a,b).** The RNA mix was then reverse transcribed using different RT enzymes and buffer conditions ($MgCl_2$ or $MnCl_2$, see also *Methods*).

We first qualitatively examined the ability of each RT enzyme to produce full length products. To this end, we used Agilent Tape Station (**Figure S2c**), which can assess cDNA product lengths independently from potential fragmentations that may occur during the library preparation and/or sequencing steps. This initial evaluation revealed that cDNA patterns strongly varied depending on the RT enzyme and buffer used. On one side, the use of manganese ($MnCl_2$) buffer conditions consistently decreased the median cDNA length compared to the use of magnesium ($MgCl_2$) buffer, possibly caused by increased RNA cleavage of Mn^{2+} ions compared to other divalent cations [26]. On the other hand, we observed that group II intron RT enzymes (Induro, TGIRT and Marathon) globally produced longer cDNA products compared to viral RT enzymes (SS2, SS3, SSIV, PS2 and Maxima), in agreement with previous works[25,27,28] .

The cDNA products were then sequenced using FIRST-seq, followed by basecalling and mapping of the reads (see *Methods*). We first assessed whether the proportion of recovered RNA biotypes varied

with RT enzyme and buffer of choice. Our analyses revealed that the use of $MnCl_2$ buffer in the RT reaction resulted in an increased proportion of recovered tRNAs across all enzymes, except SS3 **(Figure 2b,** see also **Table S4)**, suggesting that $MnCl_2$ buffer decreased the RT drop-off rates at WC-disrupting sites[28,29]. We then examined read lengths across RT enzymes and buffers, finding that observed read lengths were shorter than their corresponding cDNA products measured through TapeStation analysis **(Figure S2c)**, confirming that significant fragmentation occurs during the library preparation and sequencing steps. Despite this fragmentation, we found that several RT enzymes produced full-length reads from highly-modified rRNA molecules, with a higher proportion of full-length reads observed when using $MnCl_2$ buffer, especially when using viral RT enzymes (**Figure S3**). Finally, we assessed the sequence identity between the reference sequence and the mapped reads was consistent across RT enzymes, finding a higher sequence identity under $MgCl_2$ buffer conditions (93.4-94.3%) compared to $MnCl_2$ buffer conditions (91.7-93.4%) (**Figure S2e)**, likely attributable to an increased frequency of mismatch errors at modified positions during reverse transcription.

We then evaluated the performance of each RT enzyme with regards to its ability to introduce basecalling 'errors' at known RNA-modified sites, as this feature would be highly preferable to maximize the accuracy of single molecule RNA structure maps. To this end, we examined known RNA-modified sites in tRNAs and rRNAs, finding that $MnCl_2$ buffer enhanced processivity at WC-disrupting modifications, with increased mismatch rates and decreased RT drop-offs **(Figure 2c-d,** see also **Figure S3** and **Table S5),** suggesting that it is a better choice for RNA structure studies relying on RT misincorporation rates. We observed that the use of $MnCl_2$ buffer also increased the basecalling error at WC-neutral modified sites; a feature that could be potentially exploited to detect RNA modifications that are otherwise cannot be detected via RT errors, potentially expanding the chemical probing reagents that can be used for RNA structure determination.

Notably, all RT enzymes exhibited reduced RT drop-off scores under $MnCl_2$ buffer conditions; however, Marathon RT still displayed very high RT drop-off rates even under $MnCl_2$ buffer conditions (**Figure 2e**, see also **Table S6**), suggesting that it is a poor choice for coupling to long-read sequencing methods. Finally, we specifically examined the mismatch error frequencies at m^1A and m^3C sites under $MgCl_2$ and $MnCl_2$ buffer conditions **(Figure 2f)**, finding that mismatch frequencies at m^1A sites were only modestly affected by RT choice and buffer conditions, whereas m^3C modified sites consistently increased their mismatch frequencies under $MnCl_2$ buffer when using group II intron reverse transcriptases. Overall, these results highlight the differential effects of RT enzyme choice in the generation of full length reads and mismatch error rates at specific modified sites, and demonstrate that the use of $MnCl_2$ buffer enhances the mismatch signatures at DMS-modified sites, regardless of the enzyme of choice.

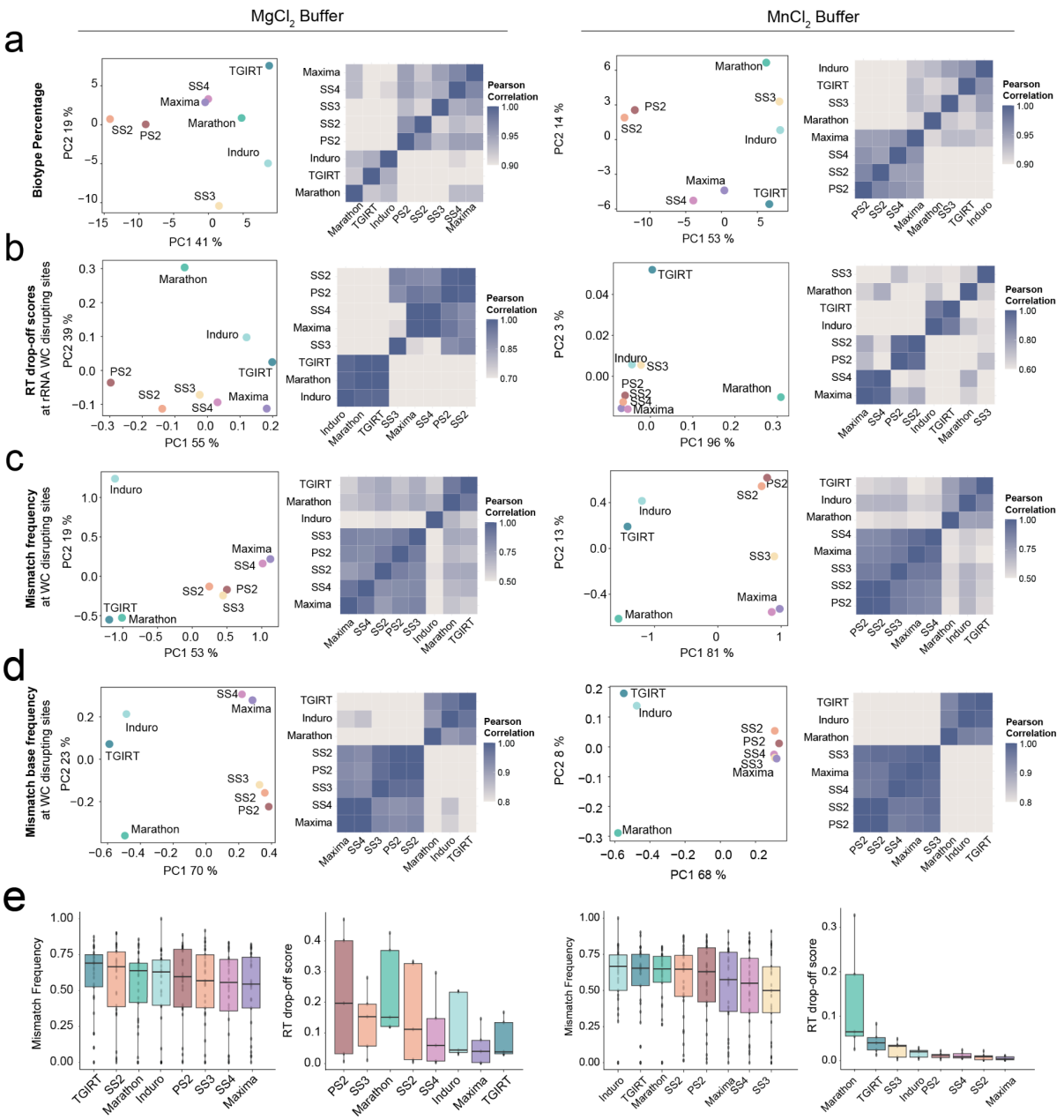


Figure 3. Signatures and performance metrics of reverse transcriptase enzymes in $MgCl_2$ and $MnCl_2$ buffers on the RNA mix sample used for benchmarking. (a) PCA plots showing the first two principal components (PC1 and PC2) based on biotype percentages from reads generated by different RT enzymes in $MgCl_2$ (left) and $MnCl_2$ (right) buffers. Each point represents an RT enzyme. Heatmap showing Pearson correlation coefficients of biotype percentages across RT enzymes in $MgCl_2$ (left) and $MnCl_2$ (right) buffers. **(b)** PCA plot showing the first two principal components (PC1 and PC2) of RT drop-off scores at rRNA Watson-Crick-disrupting (WC-disrupting) sites in $MgCl_2$ (left) and $MnCl_2$ (right) buffers. Heatmap showing Pearson correlation coefficients of RT drop-off scores at rRNA WC-disrupting sites of RT enzymes in $MgCl_2$ (left) and $MnCl_2$ (right) buffers. **(c)** PCA plots showing the first two principal components (PC1 and PC2) based on mismatch frequencies at WC disrupting sites for each RT enzyme in $MgCl_2$ (left) and $MnCl_2$ (right) buffers. Heatmap showing Pearson correlation coefficients of mismatch frequencies at WC-disrupting sites of RT enzymes in $MgCl_2$ (left) and $MnCl_2$ (right) buffers. **(d)** PCA plots showing the first two principal components (PC1 and PC2) based on mismatch base frequencies at Watson-Crick (WC) disrupting sites for each RT enzyme in $MgCl_2$ (left) and $MnCl_2$ (right) buffers. Heatmap showing Pearson correlation coefficients of mismatch base frequencies at WC-disrupting sites of RT enzymes in $MgCl_2$ (left) and $MnCl_2$ (right) buffers. **(e)** Boxplots of mismatch frequency and RT drop-off score for each RT enzyme in $MgCl_2$ (left) and $MnCl_2$ (right) buffer. For the mismatch frequency boxplots, each dot represents an m1A or m3C at rRNAs or tRNAs. For the RT drop-off score boxplots, each dot represents a WC-disrupting modification at rRNAs.

To systematically compare the performance of reverse transcriptase (RT) enzymes under $MgCl_2$ and $MnCl_2$ buffer conditions, we then performed principal component analysis (PCA) on the different features captured by FIRST-seq: i) biotype proportions, ii) mismatch frequency at modified sites, iii) mismatch base frequency at modified sites, and iv) RT drop-off rates at modified sites **(Figure 3a-e)**. Clustering by biotype revealed that viral RT enzymes (PS2 and SS2; SS4 and Maxima) and Group II intron RTs (Induro, TGIRT, and Marathon) generally clustered together, reflecting their ability to reverse transcribe molecules with similar features **(Figure 3a)**. The exception to this trend was SS3, which clustered independently from the rest of viral RT enzymes under $MgCl_2$ conditions, and clustered with Group II intron RTs under $MnCl_2$ conditions. PCA based on RT drop-off scores at WC-disrupting sites **(Figure 3b)** showed that Group II intron RTs clustered separately from M-MLV RTs, except for Marathon RT, possibly due to its high RT drop-off rates regardless of the buffer used (**Figure 3e**). PCA based on mismatch frequencies at WC-disrupting sites **(Figure 3c)** and mismatch base frequencies (**Figure 3d**) showed separate clustering of group II intron RTs and viral RT enzymes, suggesting that these two RT enzyme families produce similar error patterns upon encountering the modified site, both under $MnCl_2$ and $MgCl_2$ buffers.

Overall, our analyses show that the use of $MnCl_2$ buffer conditions generally decreases RT drop-off and increases mismatch rates (**Figure 3e**, see also **Figure S3**). Notably, group II intron RTs under $MnCl_2$ buffer conditions showed highest mismatch frequencies at WC-disrupting sites, closely followed by M-MLV RT enzymes. To our surprise, Marathon RT showed much higher RT drop-off rates than other RT enzymes, even when using $MnCl_2$ buffer conditions, making it a poor choice for long-read sequencing RNA structure probing studies. Finally, we noted that viral RT enzymes under $MnCl_2$ buffer conditions also yielded high mismatch frequency rates at modified sites while displaying minimal RT drop-offs (**Figure 3e**), suggesting that these enzymes can also be used for RNA structure studies if used with $MnCl_2$ buffer conditions. Based on these observations, we selected both Induro and SS2 RT enzymes in $MnCl_2$ buffer for DMS probing coupled to FIRST-seq experiments (DMS-FIRST-seq).

DMS-FIRST-seq accurately predicts RNA structures both *in vitro* and *in vivo*

We next assessed whether DMS-FIRST-seq would accurately recapitulate known RNA secondary structures both in DMS-probed synthetic RNAs (*B. subtilis* lysine riboswitch and *Tetrahymena* ribozyme) as well as in *in vivo* DMS-treated cells (*E. coli* and human MDA-MB231, see ***Methods***). DMS-treated samples were prepared both using DMS-FIRST-seq and commercial dRNA-seq library preparation methods, to obtain a direct comparison of the two methods on the same treated samples.

We first confirmed that DMS treatment caused increased mismatch errors, which were captured both by DMS-FIRST-seq and dRNA-seq (**Figure S4a–b**; see also **Table S7**). Errors were not randomly distributed in the case of DMS-FIRST-seq, but rather were present largely at adenosine (A) and cytosine (C) residues. By contrast, dRNA-seq showed increased errors at all four bases **(Figure S4c,d)**.

We then examined the accuracy of the methods in recapitulating known structures. To this end, DMS reactivity scores were calculated by subtracting the mismatch error rates between DMS-probed and unprobed samples on A and C positions (see also **Methods** and **Table S8**). DMS-FIRST-seq reactivities were highly consistent with known secondary structures, achieving AUC values of 0.84 and 0.86, for *B. subtilis* riboswitch and *Tetrahymena* ribozyme, respectively **(Figure S4e**). By contrast, DMS reactivity scores obtained from dRNA-seq showed lower AUC values (0.65 and 0.53, respectively) **(Figure S4e**). Visual inspection of DMS reactivity scores on the secondary structure of the *Tetrahymena* ribozyme P2–P8 domain confirmed that highly reactive bases corresponded to unpaired positions, in agreement with known folding patterns (**Figure S4f**). Together, these results confirmed that DMS-FIRST-seq outperformed dRNA-seq for accurate, base-resolution RNA structure mapping.

Next, we evaluated the performance of DMS-FIRST-seq *in vivo*. DMS incorporation was validated using LC-MS/MS, by quantifying the levels of DMS-induced modifications (m^1A, m^3C and m^7G) upon increasing DMS concentrations **(Figure 4a,b).** As previously done for IVT DMS-probed constructs, we first visualized the overall mismatch distribution along selected sequences that showed sufficient sequencing coverage, such as *Homo sapiens* RN7SL1 **(Figure 4c)** and *Escherichia coli* 16s ribosomal RNA **(Figure S5a),** finding that errors were significantly increased in DMS-treated samples compared to untreated controls **(Figure S5b-c,** see also **Tables S9-10),** as expected. Base-specific mismatch analysis revealed that errors were enriched at A and C positions, consistent with the specificity of DMS-induced modifications **(Figure 4d-e,** see also **Figure S5d-e)**, and high AUC values were obtained across distinct transcripts with known RNA structures, both for in *E. coli* and human samples, confirming the ability of FIRST-seq to capture DMS-induced errors accurately also *in vivo* (**Figure 4f,** see also **Table S11**). Notably, positions with high DMS reactivity scores mainly corresponded to unpaired positions, thus demonstrating its ability to capture RNA structural information (**Figure 4g)**.

Finally, we compared the performance of DMS-FIRST-seq depending on the enzyme of choice – i.e., Induro or SS2 RT enzymes–, by comparing the agreement between DMS reactivity and known RNA secondary structures, for each enzyme. Our results revealed that Induro RT significantly outperformed SS2 for *in vivo* RNA structure mapping at higher DMS concentrations (200mM) (**Figure S5f**). Together, our results demonstrate that DMS-FIRST-seq enables accurate and transcript-specific RNA structure mapping *in vivo,* and that Induro RT coupled with the use of Mn-based buffers constitutes an optimal solution to capture RNA structure information when using long-read sequencing technologies.

Figure 4 *(legend in next page)*

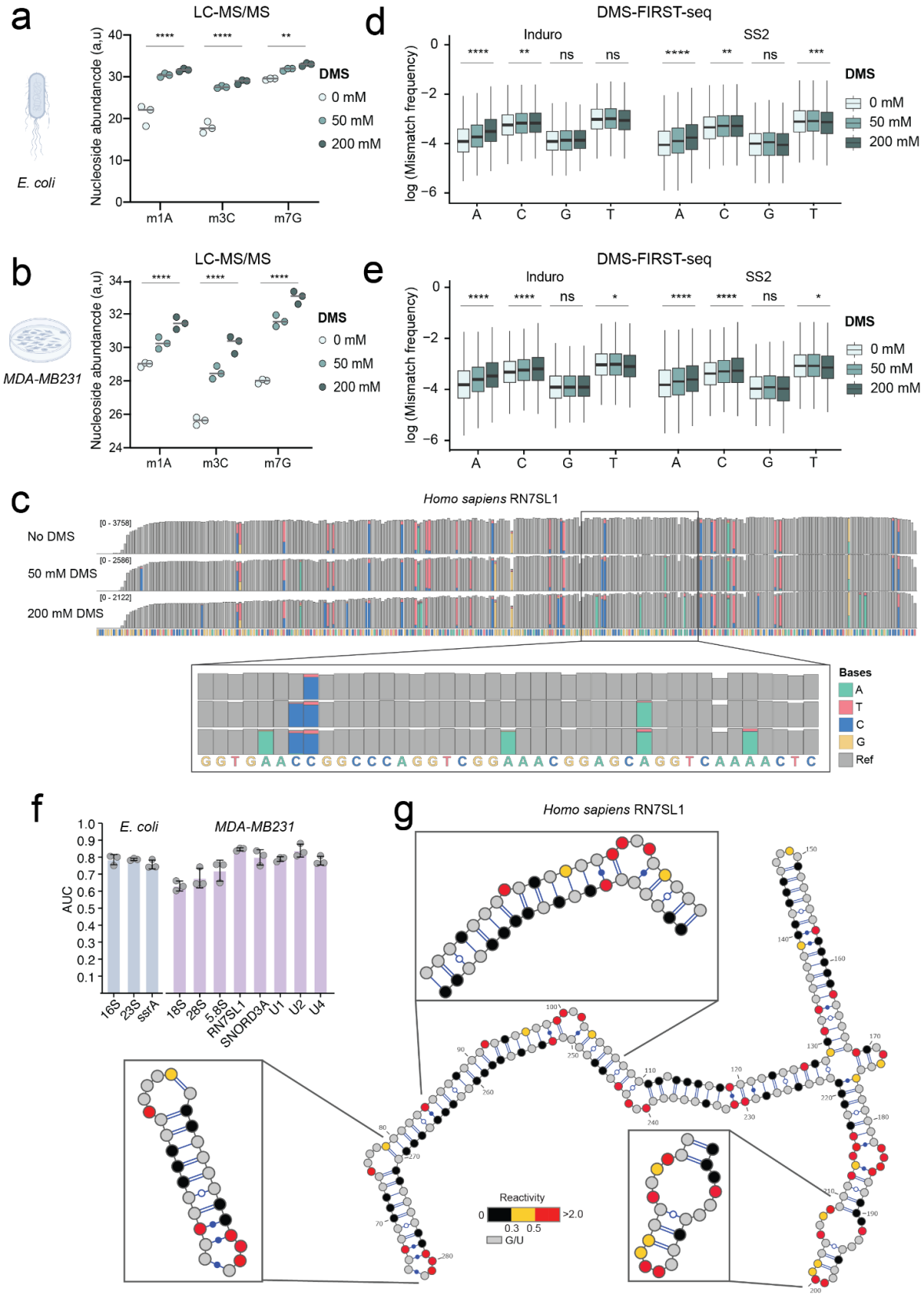


Figure 4. Structure assessment using *in vivo* DMS probing coupled to FIRST-seq in *E.coli* and *MDA-MB231* cell lines. (a-b) Dotplots showing nucleoside abundances (a.u.) of DMS induced base modifications m1A, m3C and m7G obtained using LC-MS/MS in the form of log2(Area Normalized to unmodified AUGC bases) for 0, 50 and 200 mM DMS probed *E.coli* **(a)** and MDA-MB231 **(b)** total RNAs. Each dot represents one replicate. **(c)** Integrated Genomics Viewer (IGV) coverage track snapshots show FIRST-seq data of the unprobed and DMS probed *H. sapiens* RN7SL1 RNA. Zoomed in box show a specific region within the RNA. Mismatch frequency threshold for the coverage tracks is 0.05. **(d-e)** Boxplots showing base-specific mismatch frequency of the E.coli **(d)** and MDA-MB231 **(e)** total RNA probed with different concentrations of DMS prior to sequencing with FIRST-seq using Induro and SS2 enzymes. **(f)** AUC values calculated for showing the agreement between data obtained using FIRST-seq and known structure of various RNAs in *E.coli* and *H.sapiens* (MDA-MB231) cells treated with 200 mM DMS. **(g)** Secondary structure model with overlaid base reactivities for *Homo sapiens* RN7SL1 RNA, obtained by using error differences between 200 mM DMS and 0 mM DMS treated RNAs. Structures have been generated using VARNA GUI version 1.0. Non-reactive bases (G/U) are colored in gray. For the boxplots, the box extends from the first quartile to the third quartile of the data, with a line at the median. The whiskers extend from the box to the farthest data point lying within 1.5x the interquartile range from the box. Statistical analyses were performed using the Kruskal–Wallis test ($p > 0.05$:ns, $p \leq 0.05$:*, $p \leq 0.01$:**, $p \leq 0.001$:***, $p \leq 0.0001$:****).

DISCUSSION

Nanopore direct RNA sequencing offers a simple and cost-effective platform to directly detect RNA modifications in native RNA molecules [18–20,30], including those generated by chemical probing reagents[15]. Recently, modification-aware basecalling models have been made available to detect some RNA modifications at single nucleotide and single molecule resolution[31]; however, we currently lack models for modifications induced by chemical probing reagents, such as DMS. In contrast, cDNA-based approaches benefit from generating single nucleotide resolution maps unpaired probed positions (i.e., m^1A and m^3C in case of DMS probing), while being invisible to probed paired positions that do not contain RNA structure information (i.e., m^7G modifications).

In this study, we develop FIRST-seq, a method that allows long-read transcriptome-wide nanopore cDNA sequencing by capturing only the first cDNA strand that has been synthesized, expanding the range of reverse transcriptases that can be coupled to nanopore sequencing. By overcoming the need for strand-switching, we demonstrate that FIRST-seq can be combined with any RT enzyme and buffer conditions, enabling systematic comparison of their error profiles, processivity, and susceptibility to RNA modification **(Figure 2a)**. Here, we leveraged this advantage to systematically assess the performance of RT enzymes in reverse transcribing highly modified and structured RNAs. Notably, we observed that RT choice significantly affects the biotype abundances **(Figure 2b** and **3a)**, mismatch frequencies **(Figure 2c,** and **3c-d)** and RT drop-off scores **(Figure 3b)**. Additionally, buffer composition was found to play a crucial role in RT performance, with Mg-based buffers favoring high-fidelity reverse transcription, whereas manganese-containing conditions offered increased RT processivity and mismatch error frequencies at modified sites **(Figure 2e,f)**. By coupling DMS probing with FIRST-seq (DMS-FIRST-seq), and optimal choices of RT enzymes and buffer conditions, we demonstrate that DMS reactivities correlate very well with known RNA structures **(Figure S4c-f)**, both i*n vitro* and *in vivo* DMS modified *E.coli* and human MDA-MB-231 cell line RNAs **(Figure 4a-e)**.

We should note, however, that using DMS-FIRST-seq we were unable to obtain single-molecule RNA structural maps, which was our original goal when we initiated this work. In theory, this should be achievable by increasing the frequency of DMS modifications through higher reagent concentrations. However, when employing higher DMS concentrations than those described in this work (e.g., 500 mM and above) to MDA-MB231 cells, several problems appeared. Firstly, we found a marked reduction in RNA integrity, particularly affecting large rRNAs **(Figure S6)**. Secondly, a significant decrease in the amount of recovered RNA, suggesting that highly modified transcripts are likely lost during RNA extraction. Thirdly, we observed an overall decrease in mismatch frequency at higher concentrations with increased DMS concentrations, even at positions expected to be highly reactive **(Figure S6).** This counterintuitive reduction likely reflects a technical loss of modified RNA molecules, rather than reduced chemical reactivity. Fourthly, we observed that at higher DMS concentrations, a higher proportion of 'paired' sites showed high DMS reactivity values, suggesting non-specific DMS probing. We speculate that at higher DMS concentrations there is a loss of RNA structural specificity: as transcripts become increasingly methylated, secondary structure may become destabilized, rendering previously paired regions accessible to DMS. This results in non-specific probing across the transcript, producing reactivity profiles that no longer reflect the native folding state. Together, these issues suggest that excessively high levels of DMS compromise both RNA recovery and structural fidelity, limiting the generation of accurate single-molecule RNA structure maps. Importantly, this effect may be specific to DMS. We should note that alternative probing reagents may permit higher modification densities without causing similar degradation or structural artifacts. Future work will be needed to identify and optimize such reagents for use with FIRST-seq, with the goal of enabling robust, full-length, single-molecule RNA structure mapping.

METHODS

Generation of *in vitro* transcribed RNA molecules

The list of *in vitro* transcribed constructs used in this work can be found in **Table S12**, and the corresponding fasta files can be found in https://github.com/novoalab/firstseq. *Tetrahymena* ribozyme sequence was obtained from IDT (Ref ID: 129935271). *Tetrahymena* ribozyme sequences were ordered as synthetic DNA constructs to IDT, flanked by BamHI and XhoI restriction sites and inserted into pIDTBlue plasmid. *B. subtilis* XPT_Guanine and Lysc_Lysine riboswitch sequences were ordered as synthetic DNA constructs to General Biosystems, flanked by DraI restriction sites, and inserted into a pUC57 plasmids. Sequin[32] plasmids, flanked by EcoRI-HF restriction sites (R2-63-3, R1-81-2, R1-103-1, R2-117-1) were kind gifts from Tim Mercer's lab (University of Queensland). To obtain *in vitro* transcribed RNAs, maxipreps of each of the plasmids were prepared using Qiagen Plasmid Plus Maxi Kit (Qiagen, #12963). Each construct was linearized using specific restriction enzymes. XPT_Guanine and Lysc_Lysine riboswitch sequences were linearized using DraI (NEB, R0129S), the T*etrahymena* ribozyme sequence was linearized using BamHI (NEB, R0136S) and XhoI (NEB, R0146S) and sequins sequences were linearized using EcoRI-HF (NEB, R3101S). 20 ug of each plasmid was digested overnight at 37℃ in 500 ul reaction with 10U enzyme: 1 ug plasmid ratio. Digestion product was

extracted using Phenol:Chloroform:Isoamyl alcohol (25:24:1). Digestion was confirmed with Agarose Gel Electrophoresis. The digested products of the synthetic *Tetrahymena* ribozyme, XPT_Guanine and Lysc_Lysine riboswitch sequences were *in vitro* transcribed using Ampliscribe™ T7-Flash™ Transcription Kit (Lucigen, ASF3507) following manufacturer's instructions. Sequin sequences were *in vitro* transcribed using SP6 RNA Polymerase (NEB, M0207S) following manufacturer's instructions. IVT RNA was then treated with Turbo DNase (Thermo, AM2238) (2 ul enzyme for 50 ul reaction of 200 ng/ul RNA) at 37ºC for 15 minutes, with a subsequent Zymo RNA Clean & Concentrator-5 (Zymo, R1013) cleanup, following the manufacturer's instructions. The quality of the *in vitro* transcribed (IVT) products as well as the addition of polyA tail to the synthetic constructs was assessed using Agilent 4200 Tapestation **(Figure S3a**). Concentration was determined using Qubit Fluorometric Quantitation. The purity of the IVT products was measured with NanoDrop 2000 Spectrophotometer.

Yeast culturing and total RNA extraction

Saccharomyces cerevisiae (strain BY4741) was grown at 30ºC in standard YPD medium (1% yeast extract, 2% Bacto Peptone and 2% dextrose). Cells were then quickly transferred into 50 mL pre-chilled falcon tubes, and centrifuged for 5 minutes at 3,000 g in a 4ºC pre-chilled centrifuge. Supernatant was discarded, and cells were flash frozen. Flash frozen pellets were resuspended in 700 µL TRIzol (Thermo, 15596026) with 350 µL acid washed and autoclaved glass beads (425-600 µm, Sigma G8772). The cells were disrupted using a vortex on top speed for 7 cycles of 15 seconds (the samples were chilled on ice for 30 seconds between cycles). Afterwards, the samples were incubated at room temperature for 5 minutes and 200 µL chloroform was added. After briefly vortexing the suspension, the samples were incubated for 5 minutes at room temperature. Then they were centrifuged at 14,000 g for 15 minutes at 4ºC and the upper aqueous phase was transferred to a new tube. RNA was precipitated with 2X volume Molecular Grade Absolute ethanol and 0.1X volume Sodium Acetate (3M, pH 5.2). The samples were then incubated for 1 hour at -20ºC and centrifuged at 14,000 g for 15 minutes at 4ºC. The pellet was then washed with 70% ethanol and resuspended with nuclease-free water after air drying for 5 minutes on the benchtop. Purity of the total RNA was measured with the NanoDrop 2000 Spectrophotometer. Total RNA was then treated with Turbo DNase (Thermo, AM2238) (2 ul enzyme for 50 ul reaction of 200 ng/ul RNA) at 37ºC for 15 minutes, with a subsequent RNAClean XP bead cleanup.

***In vitro* DMS probing of IVT RNAs**

1 ug IVT RNA was denatured at 95℃ for 2 mins and cooled on ice for 2 mins. Then RNA was mixed with RNA Folding Buffer (10 mM Tris pH 8.0, 100 mM NaCl, 6 mM MgCl2) up to a volume of 50 uL and incubated at 30℃ for 30 mins. For probing, 1.25 uL DMS was added to the DMS probed RNA, and 1.25 uL water was added to the unprobed RNA, and mixes were incubated at 30℃ for 5 mins. Reaction was stopped by adding 100 uL of Quenching Solution (3M B-Mercaptoethanol, 508 mM NaAC, 2ul pellet paint for each reaction), and incubated for 5 min at room temperature. RNA was precipitated by adding 300 uL absolute ethanol 100%, and resuspended in 15 uL nuclease-free water. Probed and unprobed

RNAs were then polyadenylated in vitro, to prepare them for nanopore sequencing. Poly(A) tailing was performed using *E. coli* Poly(A) Polymerase (NEB, M0276S) as per manufacturer's instructions. Poly(A) tailed RNA was recovered using Zymo RNA Clean & Concentrator-5, following the manufacturer's instructions. Concentration was determined using Qubit Fluorometric Quantitation. Purity of the IVT product was measured with a NanoDrop 2000 Spectrophotometer.

In vivo DMS probing of *Escherichia coli*

For *E. coli*, a single colony of DH5alpha bacteria was picked and inoculated in LB broth, and grown for 16h at 37°C with shaking. Bacteria were then diluted to OD600 = 0.05 and allowed to grow for an additional 2h, until OD600 ~ 0.4-0.5 (exponential phase). Dimethyl sulfate (DMS, cat. D186309, Merck), from a 1:6 dilution in ethanol (1.76 M), was directly added to the bacteria at the desired concentration. Bacteria were then shaken at 37°C for 2 min, after which DMS was immediately quenched by addition of 1 volume ice-cold DTT 1M. Bacteria were quickly pelleted by centrifugation at 17,000 g (4°C) for 1 min, the supernatant discarded, and pellets snap-frozen in liquid nitrogen.

In vivo DMS probing of MDA-MB-231

MDA-MB-231 were grown in high glucose DMEM medium, supplemented with 10% fetal bovine serum, at 37°C and 5% CO_2. After the second passage, cells at ~80% confluence were probed by direct addition of DMS, from a 1:6 dilution in ethanol (1.76 M), to the culture medium. Cells were then incubated at 37°C for 2 min, after which DMS was immediately quenched by addition of 1 volume ice-cold DTT 1M. Cells were then quickly pelleted by centrifugation at 5,000 g (4°C) for 1 min, the supernatant discarded, and pellets snap-frozen in liquid nitrogen.

RNA Isolation from *Escherichia coli and* MDA-MB-231 cells

For *E. coli*, bacterial pellets were resuspended in 62.5 µl Buffer R [20 mM Tris-HCl pH 8.0; 80 mM NaCl; 10 mM EDTA pH 8.0], supplemented with 100 µg/mL final Lysozyme (cat. L6876, Merck) and 20 U SUPERase•In™ RNase Inhibitor (cat. A2696, ThermoFisher Scientific), by thorough vortexing. After incubating samples at room temperature for 1 min, 62.5 µl Buffer L [0.5% Tween-20; 0.4% Sodium deoxycholate; 2 M NaCl; 10 mM EDTA] were added. Samples were then inverted 10 times and incubated at r.t. for 2 min, and on ice for 2 min. 1 mL ice-cold TRIzol™ Reagent (cat. 15596018, ThermoFisher Scientific) was then added, and samples were vortexed for 15 sec. RNA was extracted as per manufacturer protocol. For MDA-MB-231, cell pellets were directly resuspended in 1 mL TRIzol™ Reagent, incubated at 70°C for 5 min, and then RNA was extracted as per manufacturer protocol. Integrity of both *E. coli* and human DMS-probed total RNA was assessed using Agilent Tape Station (**Figure S7**).

Evaluation of agreement between DMS signal and structure

Agreement between DMS signal and reference structures for *in vitro* and *in vivo* RNAs was calculated using the *rf-eval* module of the RNA Framework[33]. Reference structures were obtained from the RNA Central database.

DMS reactivity score calculation

For DMS score calculation, background-subtracted raw mutation rates were normalized via the *rf-norm* module of the RNA Framework [33], using the box-plot normalization approach (parameter: *-nm 3*).

Nanopore direct RNA sequencing library preparation

The RNA libraries for direct RNA sequencing (SQK-RNA002) were prepared with the manufacturer's protocol with some adjustments. Unprobed and DMS probed RNAs were polyadenylated using *E. coli* Poly(A) Polymerase, following the commercial protocol, before starting the library prep. RNA library for direct RNA sequencing (SQK-RNA002) was prepared following the ONT Direct RNA Sequencing protocol version DRS_9080_v2_revI_14Aug2019. Briefly 500 ng total of poly(A) tailed RNAs were ligated to RT adaptors (RTA) using concentrated T4 DNA Ligase (NEB, M0202T) and was reverse transcribed using TGIRT, without the heat inactivation step. The products were purified using 1.8X Agencourt RNAClean XP beads (Thermo, NC0068576) and washed with 70% freshly prepared ethanol. Then, RMX adapter, composed of sequencing adapters with motor protein, was ligated onto the RNA:DNA hybrid, and the mix was purified using 1X Agencourt RNAClean XP beads, washing with buffer twice. The sample was then eluted in elution buffer and mixed with RNA running buffer before loading onto a primed R9.4.1 flowcell and run on a MinION sequencer with MinKNOW acquisition software version v.3.5.5.

Nanopore direct cDNA library preparation

The cDNA libraries for the direct cDNA sequencing were prepared following the direct cDNA Sequencing ONT protocol (DCB_9091_v109_revC_04Feb2019) and barcoded using the Native Barcoding Expansion kit (EXP-NBD104). Briefly, 50 ng poly(A) tailed RNA was mixed with 2.5 uL VNP oligo, 1 uL dNTPs (Thermo, R0192), and up to 11 uL nuclease-free water. After gently mixing by flicking and spinning down briefly, the mix was incubated at 65°C for 5 minutes and then placed on ice immediately. In a separate tube, 4 uL 5X RT Buffer, 1 uL RNase Inhibitor Murine (NEB, M0314S), 2 uL Strand-Switching Primer (SSP), 1 uL nuclease-free water was prepared and then two solutions were mixed. After incubating the mix at 42° C for 2 minutes, 1 uL Maxima H Minus Reverse Transcriptase (Thermo, EP0752) was added. Then the mix was incubated at 42°C for 90 minutes and inactivated by heating at 85°C for 5 minutes before moving to ice. RNAse Cocktail (Thermo Scientific, AM2286) was added to the mix in order to digest the RNA, and the mix was incubated at 37°C for 10 minutes. The reaction was then cleaned up using 0.8X AMPure XP Beads (Agencourt, A63881). Second strand synthesis was performed by mixing the 20 uL cDNA sample with 25 uL 2X LongAmp Taq Master Mix (NEB, M0287S), 2 uL PR2 Primer (PR2), and 3 uL nuclease-free water. The reaction was incubated at 94 °C for 1 min,

50 °C for 1 min, 65 °C 15 mins, and placed on ice. The reaction was then cleaned up using 0.8X AMPure XP Beads. End-repair and dA-tailing of the double-stranded DNA was achieved by mixing the 20 uL DNA sample with 30 uL nuclease-free water, 7 uL Ultra II End-prep reaction buffer and 3 uL Ultra II End-prep enzyme mix (NEB, E7546S), and incubating this mix at 20° C for 5 minutes and 65° C for 5 minutes. After incubating the reaction on ice for 1 min, it was cleaned up using 0.8X AMPure XP Beads. End-prepped DNA was then ligated to the Native Barcode (NB) for multiplexing with other libraries. The ligation was performed by mixing 22.5 uL end-prepped DNA, 2.5 uL Native Barcode (NB), 25 uL Blunt/TA Ligation Master Mix (NEB, M0367S) and incubating the mix at room temperature for 10 mins. The reaction was then cleaned up using 0.8X AMPure XP Beads. Once all the barcoded libraries were pooled after elution, the ligation of the sequencing Adapter Mix (AMX) was performed by mixing the 65 uL pooled barcoded DNAs (up to 200 fmol), 5 uL AMX, 20 uL 5X NEBNext Quick Ligation Reaction Buffer (NEB,B6058S) and 10 uL Quick T4 DNA Ligase (NEB,M2200S), and incubating the mix at room temperature for 10 mins. The reaction was cleaned up using 0.8X AMPure XP beads, using WSB Buffer for washing. The sample was then eluted in Elution Buffer (EB) and mixed with Sequencing Buffer (SQB) and Loading Beads (LB) prior to loading onto a primed R9.4.1 flowcell. Library was run on MinION flowcell with MinKNOW acquisition software version v.3.5.5.

FIRST-seq library preparation

Before starting the library preparation, 2 µL of 10 µM VNP (/5Phos/ACTTGCCTGTCGCTCTATCTTCTTTTTTTTTTTTTTTTTTTTVN) and 2 µL of 10 µM Comp_DNA (GAAGATAGAGCGACAGGCAAGTA) were mixed with 1 µL 0.1 M Tris pH 7.5, 1 µL 0.5 M NaCl, and 4 ul RNase-free water. The mix was incubated at 94°C for 1 min and the temperature was ramped down to 25°C (-0.1°C/s) in order to pre-anneal the oligos. For each reverse-transcriptase, different reverse-transcription reaction was prepared with 50 ng poly(A) tailed RNA and 1 uL pre-annealed oligo. For Superscript II (Thermo, 18064014), RNA and oligos were mixed with 1uL dNTPs and RNase-free water up to 11 uL. Mixture was heated to 65°C for 5 mins and then quickly chilled on ice. Then 4 uL 5X First-strand buffer, 1 uL 0.1 M DTT, 1 uL RNAse Inhibitor Murine, and 1 uL SuperScript II were added to the mix. Reaction was then incubated at 42°C for 60 mins and inactivated by heating at 70°C for 15 mins. For Protoscript II (NEB, M0368S), same reaction was performed with the 5X ProtoScript II buffer. For Superscript III (Thermo, 18080093), RNA and oligos were mixed with 1uL dNTPs, 4 uL 5X First-strand buffer, 1 uL 0.1 M DTT, 1 uL RNAse Inhibitor Murine, 1 uL Superscript III and RNase-free water up to 11 uL. Reaction was then incubated at 55°C for 60 mins and inactivated by heating at 70°C for 15 mins. For Superscript IV (Thermo, 18090010), RNA and oligos were mixed with 1uL dNTPs, 4 uL 5X SSIV buffer, 1 uL 0.1 M DTT, 1 uL RNAse Inhibitor Murine, 1 uL Superscript IV and RNase-free water up to 11 uL. Reaction was then incubated at 60°C for 30 mins and inactivated by heating at 80°C for 10 mins. For Maxima H Minus RT, RNA and oligos were mixed with 1 uL dNTPs, 4 uL 5X RT buffer, 1 uL RNAse Inhibitor Murine, 1 uL Maxima enzyme and RNase-free water up to 14 uL. Reaction was then incubated at 60°C for 30 mins and inactivated by heating at 85°C for 5 mins. For TGIRT (InGex), RNA and oligos were mixed with 4 uL 5X TGIRT Reaction Buffer (2.25 M NaCl, 25 mM

MgCl2, 100 mM Tris-HCl, pH 7.5), 1 uL 0.1 M DTT, 1 uL RNAse Inhibitor Murine, 1 uL TGIRT and RNase-free water up to 9.75 uL. Mix was pre-incubated at room temperature for 30 minutes, then 1.25 uL dNTP was added. Reaction was then incubated at 60°C for 60 mins and inactivated by heating at 75°C for 15 mins. For Induro RT (NEB, M0681S), RNA and oligos were mixed with 1uL dNTPs, 4 uL 5X Induro buffer, 1 uL RNAse Inhibitor Murine, 1 uL Induro and RNase-free water up to 11 uL. Reaction was then incubated at 60°C for 30 mins and inactivated by heating at 95°C for 10 mins. For Marathon RT (Kerafast), RNA and oligos were mixed with 1uL dNTPs, 10 uL 2X Marathon buffer (400 mM KCl, 4 mM MgCl2, 100 mM Tris-HCl, pH 8.3, 10 mM DTT, 40% Glycerol), 1 uL RNAse Inhibitor Murine, 1 uL Marathon and RNase-free water up to 6 uL. Reaction was then incubated at 42°C for 30 mins and inactivated by heating at 95°C for 1 mins. For the manganese buffer conditions of Protoscript II, Superscript II, Superscript III, Superscript IV, Maxima, and TGIRT, buffers disclosed by the manufacturer were prepared by replacing Magnesium with the same amount of Manganese Chloride (Thermo, AAJ63150AD). For Induro, since the manufacturer does not disclose the RT buffer composition, the TGIRT buffer (with Manganese) was used. For Marathon RT, as suggested in literature [28], 2X Marathon RT buffer was prepared, replacing Magnesium with 1 mM Manganese. Once the reverse transcription reaction was performed, RNAse Cocktail was added to the mix in order to digest the RNA, and the mix was incubated at 37°C for 10 minutes. The reaction was then cleaned up using 0.8X AMPure XP Beads. In order to be able to ligate the sequencing adapters to the first cDNA strand, 1 µL 25 µM Comp_DNA was re-annealed to the 15 µL cDNA in a tube with 2.25 µL 0.1 M Tris pH 7.5, 2.25 µL 0.5 M NaCl and 2 µL nuclease-free water. The mix was incubated at 90°C for 1 minute and the temperature was ramped down to 25 °C (-0.1°C/s) in order to anneal the complementary to the first strand cDNA. The ligation of the barcode was performed by mixing 22.5 uL end-prepped DNA, 2.5 uL Native Barcode (NB), 25 uL Blunt/TA Ligation Master Mix (NEB, M0367S) and incubating the mix at room temperature for 10 mins. The reaction was then cleaned up using 0.8X AMPure XP Beads. Once all the barcoded libraries were pooled after elution, the ligation of the sequencing Adapter Mix (AMX) was performed by mixing the 65 uL pooled barcoded DNAs (up to 200 fmol), 5 uL AMX, 20 uL 5X NEBNext Quick Ligation Reaction Buffer and 10 uL Quick T4 DNA Ligase, and incubating the mix at room temperature for 10 mins. The reaction was cleaned up using 0.8X AMPure XP beads, using WSB Buffer for washing. The sample was then eluted in Elution Buffer (EB) and mixed with Sequencing Buffer (SQB) and Loading Beads (LB) prior to loading onto a primed R9.4.1 flowcell. Library was run on MinION flowcell with MinKNOW acquisition software version v.3.5.5. A detailed step-by-step DMS-FIRST-seq protocol is provided as an additional file, both for R9 and R10 flowcells (**File S1**). We should note that DMS-FIRST-seq libraries of *in vitro* DMS modified RNAs were prepared using TGIRT in manganese buffer. *In vivo* DMS modified RNAs, Induro in manganese buffer was used (TGIRT is not commercially available any more).

Analysis of dRNA-seq datasets

For the yeast ribosomal RNA data, reads were downloaded from European Nucleotide Archive (ENA) under accession number PRJEB37798. Fast5 reads were basecalled using Guppy v6, under the RNA basecalling model, using default parameters. Prior to mapping, uridines were converted to thymidines

using the command "awk '{ if (NR%4==2) { gsub(/U/,"T",$1); print $1 } else print }' file.fastq". Reads were mapped using minimap2[34] version 2.26-r1175, with the default parameters to the *S. cerevisiae* 5s, 5.8s, 18s and 28s rRNA reference sequences.

For the yeast Nano-tRNAseq data, reads were downloaded from European Nucleotide Archive (ENA) under accession number PRJEB55684. Fast5 reads were basecalled using Guppy v6, under the RNA basecalling model, using default parameters. Prior to mapping, uridines were converted to thymidines using the command "awk '{ if (NR%4==2) { gsub(/U/,"T",$1); print $1 } else print }' file.fastq". Reads were mapped using BWA version 0.7.17-r1188, using custom parameters (bwa mem -W13 -k6 -xont2d -T20) to the *S. cerevisiae* tRNA reference sequences.

For the *in vitro* DMS treated samples, fast5 reads were basecalled using Guppy v6, under the RNA basecalling model, using default parameters. Prior to mapping, uridines were converted to thymidines using the command "awk '{if (NR%4==2) {gsub(/U/,"T",$1); print $1 } else print }' file.fastq". Reads were mapped using graphmap version v0.5.2 , with the default parameters to the in vitro RNA sequences (Tetrahymena ribozyme and lysine riboswitch RNAs).

All the mapped sam files were then converted to bam, sorted and indexed using samtools version 1.19. Mapping features such as base quality, mismatches, insertions and deletions were extracted for each position using in-house *bamtostats.sh* script. All reference files used in this work are available in the GitHub repository (https://github.com/novoalab/firstseq). Statistics for each sequencing run can be found on **Table S13**.

Annotations

The list of tRNA modifications present in *S. cerevisiae* tRNAs was obtained from MODOMICS (https://iimcb.genesilico.pl/modomics/sequences/) and was retrieved on 21 September 2021. tRNA expression estimates from ONT-based *S. cerevisiae* tRNA sequencing were obtained from Nano-tRNAseq [35], and Illumina-based *S. cerevisiae* tRNA sequencing were obtained from mim-tRNAseq [36] (GSE152621).

Analysis of dcDNA-seq datasets and FIRST-seq

For the dcDNA-seq and FIRST-seq data of the yeast total RNA runs from various reverse-transcriptases, fast5 reads were basecalled using Guppy v6, under the DNA high-accuracy (hac) basecalling model.

Then, for dcDNA-seq dataset, reads were mapped to two reference sequences using a two-step approach. Initially, reads were mapped to the long reference sequence, which includes non-tRNA references (*S. cerevisiae* non-coding RNA biotypes and in vitro transcripts) using minimap2 version 2.26-r1175, with the parameters -ax map-ont --MD -t 2. Mapped reads were filtered using samtools

version 1.19 (view -Sb -bq 59), sorted, and indexed. Unmapped reads were extracted, and the remaining reads were filtered using seqkit. The filtered reads were subsequently mapped to the short reference sequence, which includes *S. cerevisiae* tRNA references, using BWA version 0.7.17-r1188 (mem -W 13 -k 6 -x ont2d). Mapped reads were sorted and indexed using samtools. Both long and short mapped datasets were merged using Picard's MergeSamFiles tool, with MSD=true and ASSUME_SORTED=false parameters. The final merged BAM files were indexed using samtools.

For FIRST-seq datasets, the same pipeline was performed, with an addition of filtering out for the reverse strand reads with the option of -f 0x10 in samtools view command.

To analyze reverse transcription (RT) stops, read start and end positions were extracted from the BAM files using the in-house read_ends.sh script. Mapping features such as base quality, mismatches, insertions and deletions were extracted for each position using in-house *bamtostats.sh* script. All reference files used in this work are available in the GitHub repository (https://github.com/novoalab/firstseq). Statistics for each sequencing run can be found on **Table S13**.

Analysis of DMS-FIRST-seq datasets

Fast5 reads of DMS-FIRST-seq experiments were basecalled using Guppy v6, under the DNA high-accuracy (hac) basecalling model. For the DMS-FIRST-seq data of the *in vitro* RNAs, reads were mapped to the in vitro RNA sequences (Tetrahymena ribozyme and lysine riboswitch RNAs), using minimap2 version 2.26-r1175, with the parameters -ax map-ont --MD -t 2. All the mapped sam files were then converted to bam, sorted and indexed using samtools version 1.19. Mapping features such as base quality, mismatches, insertions and deletions were extracted for each position using in-house *bamtostats.sh* script. For the DMS-FIRST-seq data of the *in vivo E.coli* and MDA-MB-231 samples, reads were mapped to selected structured RNA references for each organism, using minimap2 version 2.26-r1175, with the parameters -ax map-ont --MD -t 2. All the mapped sam files were then converted to bam, sorted and indexed using samtools version 1.19. Mapping features such as base quality, mismatches, insertions and deletions were extracted for each position using in-house *bamtostats.sh* script. All reference files used in this work are available in the GitHub repository (https://github.com/novoalab/firstseq). Statistics for each sequencing run can be found on **Table S13**.

LC-MS/MS sample preparation and data analysis

Samples were digested with 1 µl of the Nucleoside Digestion Mix (New England BioLabs, M0649S) and the mixture was incubated at 37 ºC for 1 h. Samples (20ng) were analyzed using an Orbitrap Eclipse Tribrid mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) coupled to an EASY-nLC 1200 (Thermo Fisher Scientific (Proxeon), Odense, Denmark). Ribonucleosides were loaded directly onto the analytical column and were separated by reversed-phase chromatography using a 50-cm column with an inner diameter of 75 µm, packed with 2 µm C18 particles (Thermo Fisher Scientific, ES903). Chromatographic gradients started at 93% buffer A and 3% buffer B with a flow rate of 250

nl/min for 5 minutes and gradually increased to 30% buffer B and 70% buffer A in 20 min. After each analysis, the column was washed for 10min with 0% buffer A and 100% buffer B. Buffer A: 0.1% formic acid in water. Buffer B: 0.1% formic acid in 80% acetonitrile. The mass spectrometer was operated in positive ionization mode with nanospray voltage set at 2.4 kV and source temperature at 305°C. For Parallel Reaction Monitoring (PRM) method the quadrupole isolation window was set to 0.7 m/z, and MSMS scans were collected over a mass range of m/z 50-450, with detection in the Orbitrap at resolution of 120,000. MSMS fragmentation of defined masses and schedule retention time **(Table S14**) was performed using HCD at NCE 20 (except stated differently, **Table S14),** the auto gain control (AGC) was set to "Standard" and a maximum injection time of 246 ms was used. In each PRM cycle, one full MS scan at resolution of 120,000 was acquired over a mass range of m/z 220-700 with detection in the Orbitrap mass analyzer. Auto gain control (AGC) was set to 1e5 and the maximum injection time was set to 50 ms. Serial dilutions were prepared using commercial pure ribonucleosides (0.005-150 pg, Carbosynth, Toronto Research Chemicals) in order to establish the linear range of quantification and the limit of detection of each compound [37] A mix of commercial ribonucleosides was injected before and after each batch of samples to asses instrument stability and to be used as external standard to calibrate the retention time of each ribonucleoside.

Acquired data were analyzed with the Skyline-daily software (v24.1.1.284) and extracted precursor areas of the ribonucleosides were used for quantification. The raw proteomics data have been deposited to the MetaboLights [38] repository with the dataset identifier REQ20250127208298.

DATA AVAILABILITY

Data generated in this study, including basecalled FASTQ and processed data, has been deposited in GEO, under accession GSE288618.

CODE AVAILABILITY

Code to analyze DMS-probed nanopore cDNA data has been deposited in GitHub (https://github.com/novoalab/firstseq).

AUTHOR CONTRIBUTIONS

OB performed most wet lab experiments and bioinformatic analyses included in this work, and made all figures. GD contributed with validation experiments of DMS-probed *in vivo* RNA structures. JSM contributed with supervision of this work. IV and DI provided vivo DMS-probed samples from *E. coli* and human cell lines. EMN conceived and supervised the project. OB and EMN wrote the manuscript, with contributions from all authors.

ACKNOWLEDGEMENTS

We thank all the members of the Novoa lab for their insightful discussions. We are grateful to the CRG Core Technologies Programme for their support and assistance in this work

FUNDING

GD was supported by a Marie Sklodowska-Curie fellowship ("ROPES"), under agreement No 956810 (GD). This work was supported by the Spanish Ministry of Science and Innovation (PID2021-128193NB-I00 and MCIN/AEI/10.13039/501100011033/ FEDER, UE); the European Research Council (ERC Starting Grant 101042103 and ERC Proof-of-Concept 101187456) and the Catalan Agency for Research and Universities (SGR-2021-01301). We acknowledge support of the Spanish Ministry of Science and Innovation through the Centro de Excelencia Severo Ochoa (CEX2020-001049-S, MCIN/AEI /10.13039/501100011033), and the Generalitat de Catalunya through the CERCA programme and to the EMBL partnership. JSM is supported by UNSW Sydney grant RG193211. The mass spectrometry analyses were performed in the CRG/UPF Proteomics Unit which is part of the Spanish National Infrastructure for Omics Technologies (ICTS OmicsTech).

COMPETING INTERESTS

EMN is a member of the Scientific Advisory Board of IMMAGINA Biotech. JSM is a member of the Scientific Advisory Board of NextRNA Therapeutics Inc. OB, GD and EMN have received travel bursaries from ONT to present their work at conferences.

REFERENCES

1. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2020).
2. Wilkinson, E., Cui, Y.-H. & He, Y.-Y. Roles of RNA Modifications in Diverse Cellular Functions. *Frontiers in Cell and Developmental Biology* **10**, (2022).
3. Ouyang, J. *et al.* Long non-coding RNAs are involved in alternative splicing and promote cancer progression. *Br. J. Cancer* **126**, 1113–1124 (2021).
4. Mattick, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat. Rev. Mol. Cell Biol.* **24**, 430–447 (2023).
5. Spitale, R. C. & Incarnato, D. Probing the dynamic RNA structurome and its functions. *Nat. Rev. Genet.* **24**, 178–196 (2022).
6. Tijerina, P., Mohr, S. & Russell, R. DMS footprinting of structured RNAs and RNA-protein complexes. *Nat. Protoc.* **2**, 2608–2623 (2007).

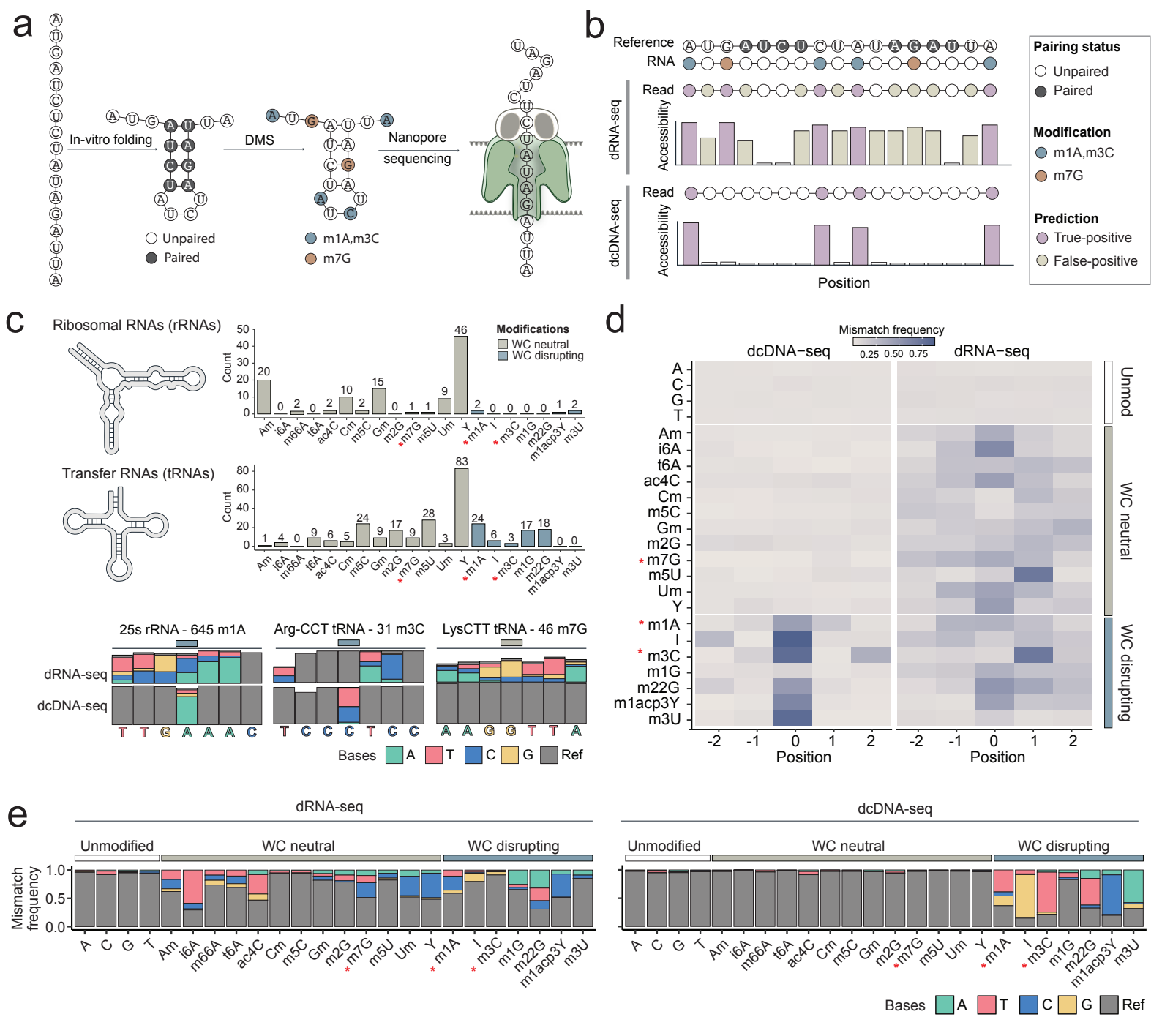
7. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 97–102 (2009).
8. Ehresmann, C. *et al.* Probing the structure of RNAs in solution. *Nucleic Acids Res.* **15**, 9109–9128 (1987).
9. Zubradt, M. *et al.* DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Methods* **14**, 75–82 (2017).
10. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2014).
11. Smola, M. J. & Weeks, K. M. In-cell RNA structure probing with SHAPE-MaP. *Nat. Protoc.* **13**, 1181–1195 (2018).
12. Strobel, E. J., Yu, A. M. & Lucks, J. B. High-throughput determination of RNA structures. *Nat. Rev. Genet.* **19**, 615–634 (2018).
13. Yang, M. *et al.* In vivo single-molecule analysis reveals COOLAIR RNA structural diversity. *Nature* **609**, 394–399 (2022).
14. Aw, J. G. A. *et al.* Determination of isoform-specific RNA structure with nanopore long reads. *Nat. Biotechnol.* **39**, 336–346 (2020).
15. Stephenson, W. *et al.* Direct detection of RNA modifications and structure using single-molecule nanopore sequencing. *Cell genomics* **2**, 100097 (2022).
16. Bohn, P., Gribling-Burrer, A.-S., Ambi, U. B. & Smyth, R. P. Nano-DMS-MaP allows isoform-specific RNA structure determination. *Nat. Methods* **20**, 849–859 (2023).
17. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
18. Liu, H. *et al.* Accurate detection of m6A RNA modifications in native RNA sequences. *Nat. Commun.* **10**, 4079 (2019).
19. Begik, O. *et al.* Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat. Biotech.* (2021) doi:10.1038/s41587-021-00915-6.
20. Leger, A. *et al.* RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat. Commun.* **12**, 1–17 (2021).
21. Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R. & Akeson, M. Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLoS One*

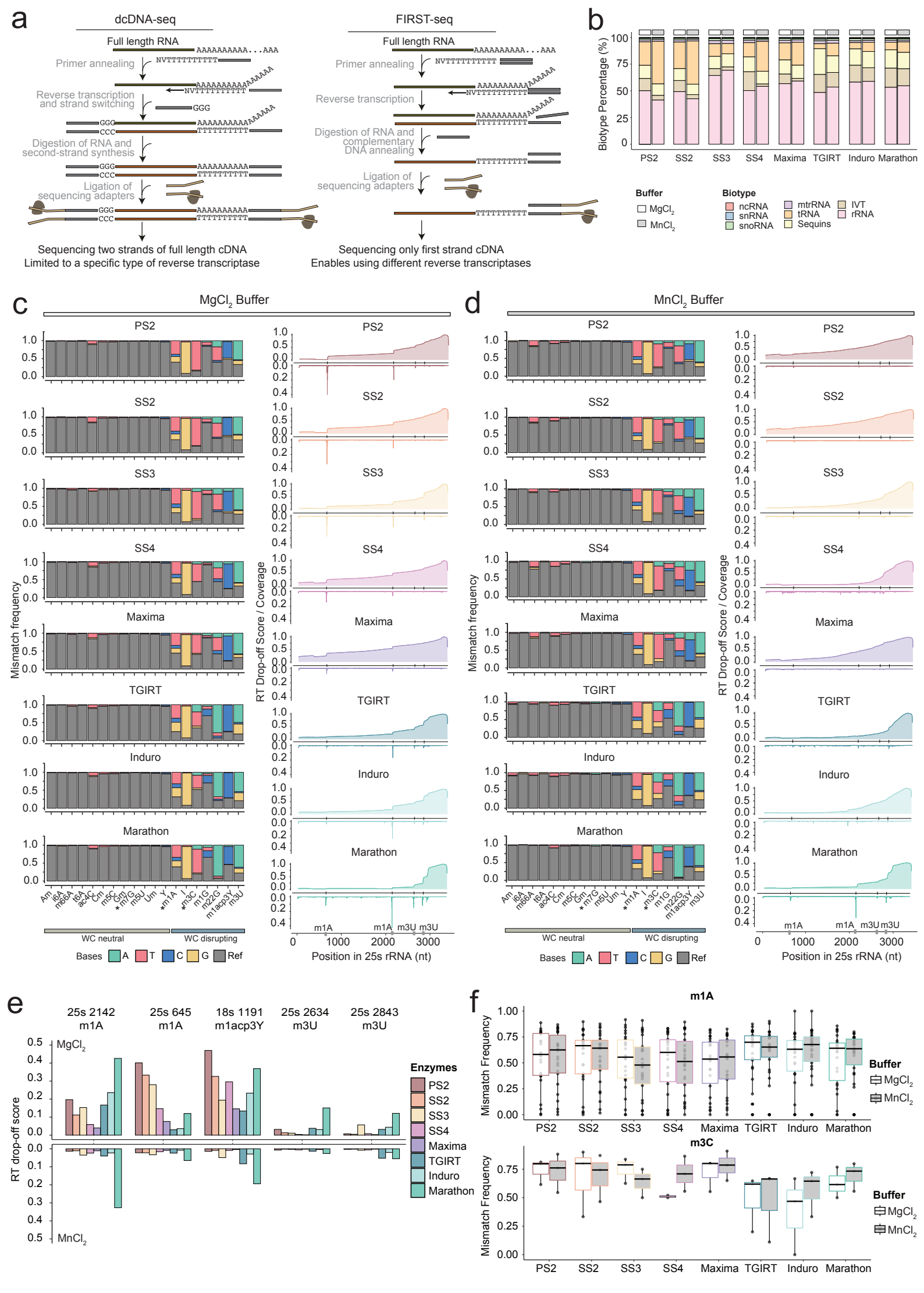
14, e0216709 (2019).

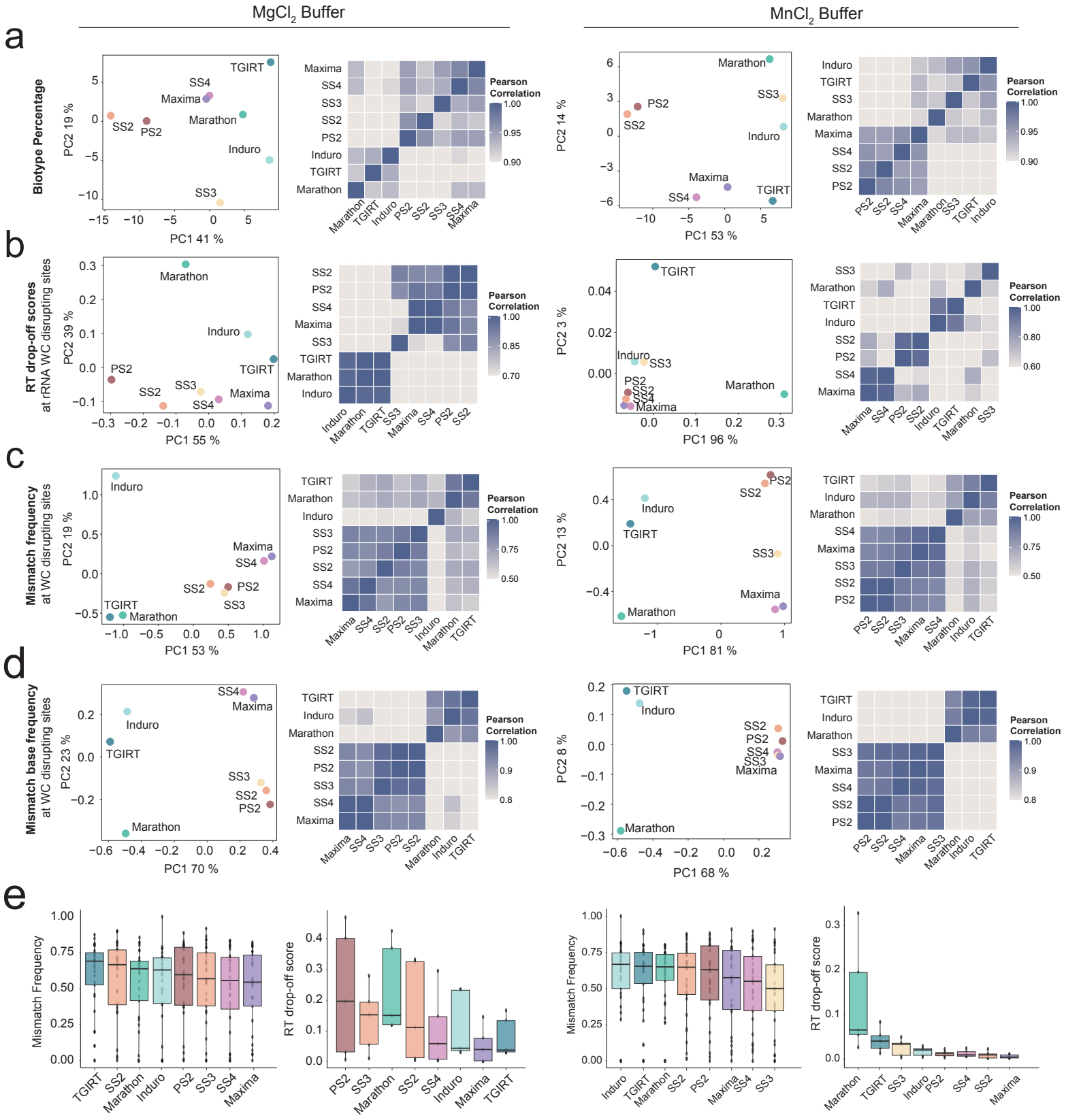
22. Soneson, C. *et al.* A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.* **10**, 3359 (2019).
23. Ura, H., Togi, S. & Niida, Y. Poly(A) capture full length cDNA sequencing improves the accuracy and detection ability of transcript quantification and alternative splicing events. *Sci. Rep.* **12**, 10599 (2022).
24. Byrne, A. *et al.* Single-cell long-read targeted sequencing reveals transcriptional variation in ovarian cancer. *Nat. Commun.* **15**, 6916 (2024).
25. 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).
26. Wrzesinski, J., Michałowski, D., Ciesiołka, J. & Krzyzosiak, W. J. Specific RNA cleavages induced by manganese ions. *FEBS Lett.* **374**, 62–68 (1995).
27. Qin, Y. *et al.* High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. *RNA* **22**, 111–128 (2016).
28. Guo, L.-T. *et al.* Sequencing and Structure Probing of Long RNAs Using MarathonRT: A Next-Generation Reverse Transcriptase. *J. Mol. Biol.* **432**, 3338–3352 (2020).
29. Kristen, M. *et al.* Manganese ions individually alter the reverse transcription signature of modified ribonucleosides. *Genes (Basel)* **11**, 950 (2020).
30. Begik, O., Mattick, J. S. & Novoa, E. M. Exploring the epitranscriptome by native RNA sequencing. *RNA* **28**, 1430–1439 (2022).
31. Cruciani, S. *et al.* De novo basecalling of RNA modifications at single molecule and nucleotide resolution. *Genome Biol.* **26**, 38 (2025).
32. Hardwick, S. A. *et al.* Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods* **13**, 792–798 (2016).
33. Incarnato, D., Morandi, E., Simon, L. M. & Oliviero, S. RNA Framework: an all-in-one toolkit for the analysis of RNA structures and post-transcriptional modifications. *Nucl. Acids Res.* **46**, e97 (2018).
34. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
35. Lucas, M. C. *et al.* Quantitative analysis of tRNA abundance and modifications by nanopore RNA

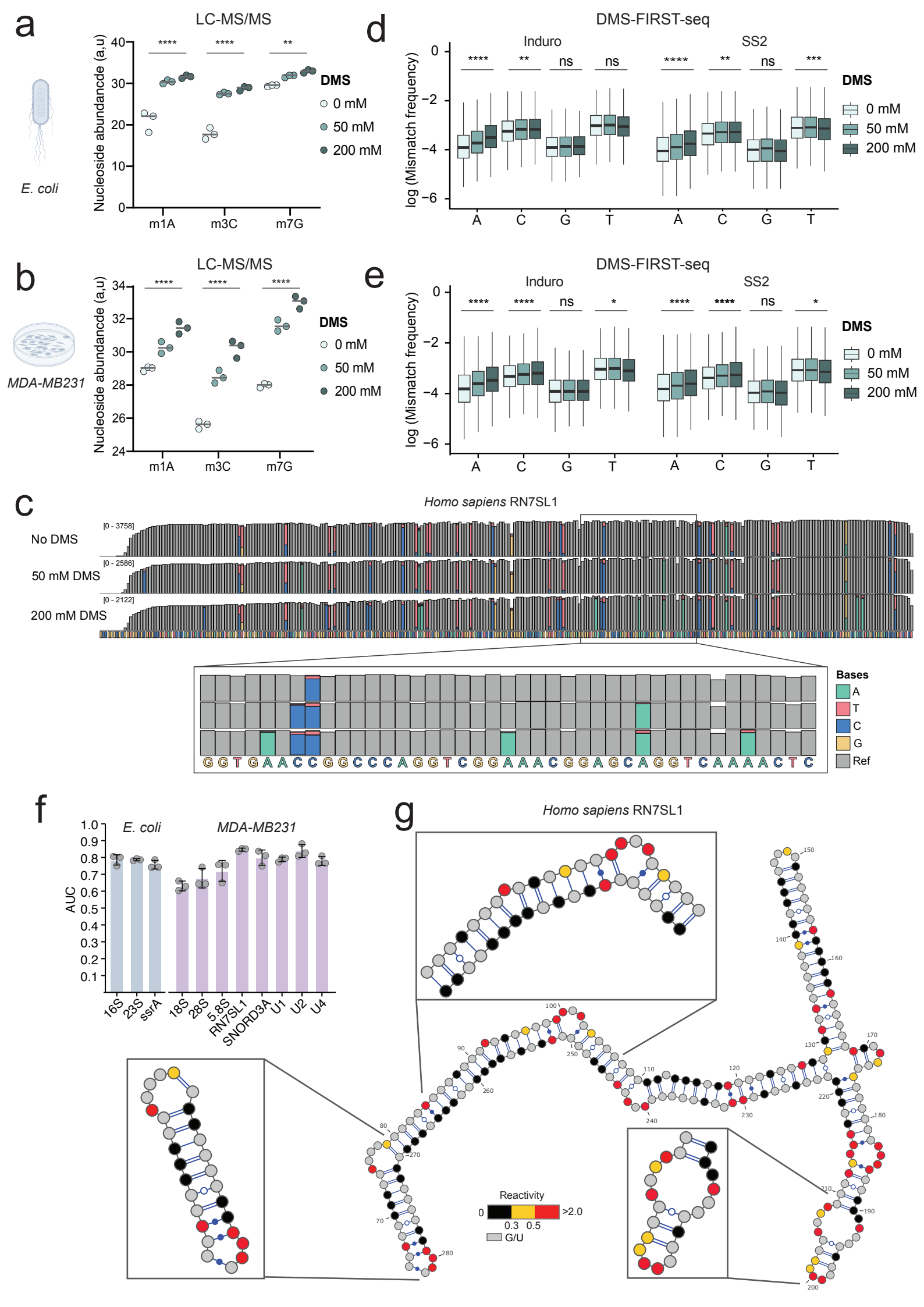
sequencing. *Nat. Biotech.* **42**, 72–86 (2023).

36. Behrens, A., Rodschinka, G. & Nedialkova, D. D. High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. *Mol Cell* **81**, 1802–1815.e7 (2021).
37. Sugimoto, N. *Handbook of Chemical Biology of Nucleic Acids*. (Springer Nature, 2023).
38. Yurekten, O. *et al.* MetaboLights: open data repository for metabolomics. *Nucl. Acids Res* **52**, D640–D646 (2024).


a
In-vitro folding
DMS
Nanopore sequencing
Unpaired
Paired
m1A,m3C
m7G
b
Reference
RNA
Read
Accessibility
dRNA-seq
dcDNA-seq
Position
Pairing status
Unpaired
Paired
Modification
m1A,m3C
m7G
Prediction
True-positive
False-positive
c
Ribosomal RNAs (rRNAs)
Transfer RNAs (tRNAs)
Count
Modifications
WC neutral
WC disrupting
Am i6A m6A t6A ac4C Cm m5C Gm m2G *m7G m5U Um Y *m1A I *m3C m1G m22G m1acp3Y m3U
25s rRNA - 645 m1A
Arg-CCT tRNA - 31 m3C
LysCTT tRNA - 46 m7G
dRNA-seq
dcDNA-seq
Bases A T C G Ref
d
Mismatch frequency
0.25 0.50 0.75
dcDNA−seq
dRNA−seq
Unmod
WC neutral
WC disrupting
Position
e
dRNA-seq
dcDNA-seq
Unmodified
WC neutral
WC disrupting
Mismatch frequency
Bases A T C G Ref


a
dcDNA-seq
Full length RNA
Primer annealing
Reverse transcription and strand switching
Digestion of RNA and second-strand synthesis
Ligation of sequencing adapters
Sequencing two strands of full length cDNA
Limited to a specific type of reverse transcriptase
FIRST-seq
Full length RNA
Primer annealing
Reverse transcription
Digestion of RNA and complementary DNA annealing
Ligation of sequencing adapters
Sequencing only first strand cDNA
Enables using different reverse transcriptases
b
Biotype Percentage (%)
PS2
SS2
SS3
SS4
Maxima
TGIRT
Induro
Marathon
Buffer
MgCl2
MnCl2
Biotype
ncRNA
snRNA
snoRNA
mtrRNA
tRNA
Sequins
IVT
rRNA
c
MgCl2 Buffer
Mismatch frequency
RT Drop-off Score / Coverage
WC neutral
WC disrupting
Bases
A
T
C
G
Ref
Position in 25s rRNA (nt)
d
MnCl2 Buffer
e
25s 2142 m1A
25s 645 m1A
18s 1191 m1acp3Y
25s 2634 m3U
25s 2843 m3U
RT drop-off score
Enzymes
f
m1A
m3C
Mismatch Frequency
Buffer


MgCl2 Buffer
MnCl2 Buffer
a
Biotype Percentage
PC1 41 %
PC2 19 %
PC1 53 %
PC2 14 %
Pearson Correlation
b
RT drop-off scores at rRNA WC disrupting sites
PC1 55 %
PC2 39 %
PC1 96 %
PC2 3 %
c
Mismatch frequency at WC disrupting sites
PC1 53 %
PC2 19 %
PC1 81 %
PC2 13 %
d
Mismatch base frequency at WC disrupting sites
PC1 70 %
PC2 23 %
PC1 68 %
PC2 8 %
e
Mismatch Frequency
RT drop-off score


a
LC-MS/MS
E. coli
Nucleoside abundancde (a,u)
m1A
m3C
m7G
DMS
0 mM
50 mM
200 mM
b
LC-MS/MS
MDA-MB231
d
DMS-FIRST-seq
Induro
SS2
log (Mismatch frequency)
A
C
G
T
e
c
Homo sapiens RN7SL1
No DMS
50 mM DMS
200 mM DMS
Bases
Ref
f
E. coli
MDA-MB231
AUC
16S
23S
ssrA
18S
28S
5.8S
RN7SL1
SNORD3A
U1
U2
U4
g
Homo sapiens RN7SL1
Reactivity
G/U