Best practices for genome-wide RNA structure analysis: combination of mutational profiles and drop-off information

Eva Maria Novoa[1,2,3,4,*,#], Jean-Denis Beaudoin[5,*], Antonio J Giraldez[5,6], John S Mattick[3,4] and Manolis Kellis[1,2,#]

[1] *Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 02139 Cambridge, MA, USA.*

[2] *Broad Institute of MIT and Harvard, 02139 Cambridge, MA, USA.*

[3] *Garvan Institute of Medical Research, Darlinghurst, 2010, NSW, Australia.*

[4] *University of New South Wales Sydney, Kensington NSW 2052, Australia.*

[5]*Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06510, USA*

[6]*Yale Stem Cell Center, Yale University School of Medicine, New Haven, Connecticut 06520, USA*

**co-first authors*

[#] Correspondence to: Eva Maria Novoa (e.novoa@garvan.org.au) and Manolis Kellis (manoli@mit.edu)

ABSTRACT

Genome-wide RNA structure maps have recently become available through the coupling of *in vivo* chemical probing reagents with next-generation sequencing. Initial analyses relied on the identification of truncated reverse transcription reads to identify the chemically modified nucleotides, but recent studies have shown that mutational signatures can also be used. While these two methods have been employed interchangeably, here we show that they actually provide complementary information. Consequently, analyses using exclusively one of the two methodologies may disregard a significant portion of the structural information. We find that the identity and sequence environment of the modified nucleotide greatly affects the odds of introducing a mismatch or causing reverse transcriptase drop-off. Finally, we identify specific mismatch signatures generated by dimethyl sulfate probing that can be used to remove false positives typically produced in RNA structurome analyses, and how these signatures vary depending on the reverse transcription enzyme used.

INTRODUCTION

In the last decades, it has become clear that RNAs are not simply intermediaries between DNA and protein, but are in fact functional molecules capable of regulating central cellular and developmental processes, such as genome organization and gene expression, and comprise the bulk of human genomic programming [1-4]. Because RNA is a single-stranded molecule, it tends to fold back on itself, forming stable secondary and tertiary structures by internal base pairing and other interactions. RNA structure plays an essential role in determining the function and dynamics of these molecules, and can vary depending on environmental conditions [5, 6]. Thus, accurate genome-wide RNA structural maps can allow for better understanding of the complexity, function, and regulation of the transcriptome [7, 8].

Dimethyl sulphate (DMS) and 2'hydroxyl acylation and primer extension (SHAPE) reagents have traditionally been used to obtain experimental measurements of RNA structure, providing information on base-pairing and tertiary interactions of the RNA molecules [9-11]. These chemicals show selective reactivity toward unpaired RNA bases. Until recently, limitations of probing reagents as well as sequencing and informatics had restricted structural profiling analyses to a few *in vitro* folded RNAs. In recent years, however, DMS and SHAPE chemical labeling have been coupled to next-generation sequencing [12-18], providing genome-wide RNA structure maps that have provided substantial information regarding the dynamics of RNA structures in a variety of cellular contexts [12, 15].

Initial attempts to analyze genome-wide RNA structure data relied on the truncation of reverse transcription upon reaching a nucleotide that has been chemically modified by probing reagents [13, 15, 16, 19, 20]. However, these methods present several limitations and caveats, such as the presence of naturally occurring modified nucleotides and the addition of untemplated nucleotides by the reverse transcriptase (RT) [21], hindering the correct annotation of the modified base. At least 8% of the annotated positions constitute false positives, as 8% of annotated read ends of DMS-Seq analyses are typically mapped onto G and T bases [15], while only A and C bases should be identified by this technique. Moreover, RNA ligases commonly used for adapter ligation have distinct sequence preferences for the ends being ligated, thereby biasing the representation of the read ends in the resulting libraries [22].

To overcome these limitations, several groups have used the increased mismatch rates that occur at DMS- and SHAPE-modified nucleotides [23], known as mutational profiling (MaP),

giving rise to SHAPE-MaP [24] and DMS-MaPSeq [25], respectively. However, mutations occur at low frequency, and therefore, they can only be identified in transcripts with very high read coverage. Consequently, these methods are generally limited to the analysis of individual RNA transcripts [23, 24, 26], smaller genomes such as the HIV-1 RNA genome [23], or to highly expressed genes in more complex genomes [25].

Here we have analyzed transcriptome-wide DMS-seq datasets of *in vivo* probed zebrafish embryos (see Methods) - including well-characterized RNA structures as spike-ins -, and have compared the quantification of DMS modifications using either the reverse transcription truncation signals or MaP. Previous studies have employed either reverse transcription truncation signals [15] or MaP [25], with the underlying assumption that results will largely overlap. Contrary to expectations, we find a low correlation between the two methodologies, despite correct identification of unpaired nucleotides by each methodology. Furthermore, the probability of driving a mismatch or a RT-stop signal upon reverse transcription is dependent on both the identity of the modified nucleotide as well as on the sequence context. Considering our findings, we suggest that RNA structural analyses based on DMS or SHAPE probing should combine both RT truncation signals and MaP to best capture structural information.

RESULTS

Mutational profiling and RT drop-off analyses contain non-overlapping complementary information

DMS has traditionally been used to probe RNA structure *in vivo*. It reacts with the *N1* of adenosines, *N3* of cytosines and *N7* of guanosines in single stranded (ss) RNA, resulting in chemically modified nucleosides 1-methyladenosine (m^1A), 3-methylcytosine (m^3C) and 7-methylguanosine (m^7G), respectively. Reverse transcription is typically blind to m^7G modifications, as they do not affect the hydrogen bonds involved in Watson-Crick base pairing. In contrast, methylations occurring in the *N3* moiety of cytosine (m^3C) and *N1* of adenosines (m^1A) affect the Watson-Crick base-pairing (**Figure 1A**), and can therefore cause RT drop-off [27] as well as increased error rates ('mismatch' patterns) upon reverse transcription [28, 29] (**Figure 1B**).

To compare the set of DMS-modified positions identified using either the reverse transcription truncation methodology (RT drop-off) or the mutational profiling methodology (**Figure 1B**), we first examined the RT-stop signals and mutational profiles of two spike-ins

with known RNA structure—the *Tetrahymena* ribozyme (Rz) [30] and the tRNA spinach cassette (Spi) [31]. Read coverage for these two molecules was extremely high (~50,000X), facilitating the identification of low frequency mismatch positions observed upon DMS probing with high confidence. We find that both methodologies quantify the level of per-nucleotide DMS modification with high replicability in both spike-ins (Pearson's r=0.92-0.99) (**Figure 1C**). However, when we compared the individual positions identified by each methodology, we found only a partial overlap between the two methods (**Figures 1D and 1E**), suggesting that mutational profiling and RT truncation methods are capturing non-overlapping sets of DMS-modified nucleotides.

Direct comparison of the identified positions shows that RT drop-off methodologies are unable to capture DMS-modified positions at the 5' and 3' end of the molecule (**Figures 1D and 1E**), which is an intrinsic limitation of the RT-stop methodology, due to the size selection step. In contrast, mutational profiling methods are capable of identifying DMS-modified positions in these regions. When comparing the overlap of positions identified by each methodology, we find that 40% of the positions identified using mutational profiling are not identified by RT-stop methodologies in transcriptome-wide analyses (**Figure 2A**). Similarly, in the spike-ins, 11-26% of mismatched positions are not identified by RT methodologies (**Figure 2B**).

The number of positions identified by the RT-stop signal methodology is ~2-fold higher than the mutational profiling methodology in the spike-ins (**Figure 2B**), and up to 15-fold higher when analyzing transcriptome-wide datasets (**Figure 2A**). This difference is partly due to the size-selection step in the library preparation, which enriches the sample in RT truncated reads, as well as on the coverage requirement of 50 reads/nucleotide that was used to identify mismatched positions (see Methods), which greatly limits the set of predicted mismatches to highly expressed genes.

Both methodologies correctly identify unpaired nucleotides

To ascertain whether both methodologies correctly identified unpaired positions, we superimposed the predicted mismatch and RT-stop positions onto the known secondary structure of the *Tetrahymena* ribozyme, finding that both methods correctly identify unpaired positions (**Figure 2C** and **S1**). We find that some unpaired nucleotides are preferentially identified using the RT truncation methodology (green), whilst others are preferentially identified using mutational profiling (red), further supporting the idea that the two methodologies capture non-overlapping information.

Intriguingly, most of the 'preferentially identified' nucleotides by mutational profiling appeared to be Cs, whereas most of the 'preferentially identified' nucleotides by the RT truncation method tended to be As (**Figure 2C**). Based on this observation, we hypothesized that the underlying reference nucleotide might affect the fate of the reverse transcriptase, i.e. the reverse transcriptase might preferentially incorporate mutations or drop-off depending on the identity of the modified-nucleotide. Then, we compared the signal from both methodologies at individual As and Cs (**Figure 2D**), to see if there were quantitative differences between the two. We find that, both in the spike-ins and transcriptome-wide analyses, positions predominantly identified by the mutational profiling methodology were largely Cs (red) whereas those predominantly identified by the RT-stop method corresponded to As (blue). Moreover, we observe no correlation between the two signals (**Figure 2D**). This observation further supports that both methodologies are capturing, at least in part, non-overlapping information.

Choice of mismatch or RT stop is dependent on the nature of the DMS-modified nucleotide

Previous genome-wide studies employing RT truncation methodologies to retrieve DMS-modified positions had reported that 68% of the modified positions had an adenosine as underlying nucleotide, while only 24% of the positions had a cytosine [15]. Consequently, DMS probing has been generally assumed to be biased towards modifying adenosines compared to cytosines [1]. Here we find that RT-stops tend to occur more frequently at adenosines (54%), in agreement with previous reports [15]. However, we observe that 66% of the mutational profiling signal arises from cytosines (**Figure 2E**). Our results contrast with the belief that DMS preferentially reacts with adenosines, and suggest that DMS does not preferentially react with one nucleotide or another, but instead, that the method of analysis largely determines which DMS-modified positions will be detected.

We then investigated whether the sequence context was affecting the outcome of the reverse transcription. We find that sequence contexts using either methodology appear to be slightly enriched in AT-rich contexts (**Figure 2F**), which may be due to biases introduced during library preparation [32]. When comparing the two methodologies, we observe no differences in the sequence context in the 5' vicinity of the modified positions. In contrast, we do observe that the sequence context found in the 3' vicinity of the modified positions is different (**Figure 2F**). At first sight, our analysis would suggest that the sequence context is playing a role in determining the fate of the reverse transcription (i.e. mismatch or RT-stop).

Unfortunately, library preparation biases, such as ligation bias, will also affect the sequence context identified using RT-stop methodologies, but not the sequence context identified using mutational profiling. Consequently, we cannot exclude the fact that the observed sequence context differences may be partially arising from library preparation biases that are unequally affecting the two methodologies.

DMS probing generates distinct mismatch signatures for m^3C and m^1A

In the last years, it has been suggested that each RNA modification may produce a different mismatch "signature", where the identity and relative proportion of the misincorporated nucleotides may provide a clue on the nature of the underlying RNA modification [28, 29]. Previous works have employed machine learning algorithms –trained with known tRNA modifications– to predict the nature of each RNA modification based on its mismatch pattern [28]. However, in our hands, mismatch patterns observed in tRNA molecules were not representative of those found in other RNA subtypes, likely due to the enriched modification environment and limited sequence diversity of tRNAs. In contrast, DMS-Seq allows for building mismatch signatures for both m^1A and m^3C, as these modifications will be in high abundance and in heterogeneous sequence contexts.

To characterize the mutational signal that is produced upon DMS modification, we first compared the mismatch frequencies across replicates using 64c zebrafish embryo DMS-Seq datasets (**Figure 3A**). For comparison, we analyzed mismatch frequencies across replicates of 64c zebrafish embryo RNA-Seq datasets (**Figure 3B**). As could be expected, we find that the majority of mismatches identified in RNA-Seq datasets show mismatch frequencies ~1, suggesting that all these positions actually correspond to single nucleotide polymorphisms (SNPs). In contrast, mismatches identified in DMS-Seq datasets display a bimodal distribution, where some have mismatch frequencies ~1 -corresponding to SNPs-, whereas a second population displays very low mismatch frequencies. By comparing the patterns of RNA-Seq and DMS-Seq datasets, it is clear that low frequency mismatches observed in DMS-Seq datasets are the ones of interest, i.e. appearing upon DMS treatment. Therefore, we discarded those positions whose mismatch frequency was greater than 0.25, and applied the same filter both to DMS-Seq and RNA-Seq datasets (**Figures 3A and 3B, right panels**). Surprisingly, there were still a significant amount of mismatched positions that were still present in RNA-Seq datasets. Indeed, most of these mismatch positions observed in RNA-Seq datasets were consistent across replicates, and were also found in DMS-Seq datasets (**Figure 3C**), suggesting that these positions may correspond to naturally occurring RNA

modifications or low frequency SNP alleles in our population of embryos, which can act as confounders in our analysis, and therefore, should be discarded from the analysis.

To identify the substitution patterns that occur at DMS-modified positions, we then subdivided the mismatched positions based on their reference nucleotide, and compared the relative frequencies of misincorporated nucleotides (**Figure 3D**). From the ternary plot representations, we observed that both A and C mismatch positions did not randomly incorporate nucleotides (which would lead to scattered accumulation of points near the triangle's center), but rather showed a biased "signature", introducing mismatched nucleotides at specific frequencies. More specifically, A-mismatch positions preferentially misincorporated G and T bases, while C-mismatch positions preferentially misincorporated T bases (**Figure S2**).

Upon examining all the mismatched positions, with no filter, we found that the majority (66%) of mismatched positions were found at A and C bases (**Figure 3D**). However, we also identified mismatches at G and U bases, which are not generated by DMS modification and should therefore not be observed. We hypothesized that, as we had seen before, these mismatches may be generated by SNPs, as well as by naturally occurring RNA modifications. We therefore discarded those positions with mismatch frequencies higher than 0.25, and removed mismatches with insufficient coverage or with very low frequency of mismatches, which are likely PCR/sequencing artifacts (see Methods). Applying these filters removed most of the mismatches observed at G and T bases (**Figure 3E**), supporting the validity of pairing RNA-Seq with DMS-Seq for mutational profile analysis of RNA structure datasets. The inclusion of this filtering step led to 89% of the filtered mismatches at A and C bases (**Figure 3E**). Furthermore, the mutational signature observed in A and C was enhanced, as could be expected from the mutational profile that has been generated by a single RNA modification (m^1A and m^3C, respectively), thus supporting our filtering steps as means to increase the signal-to-noise ratio, even though some true positives may be lost in the process. Thus, we conclude that this filtering step, although stringent, is essential for increasing the signal-to-noise ratio, and allows for correct identification of DMS-specific mutational profiles (**Figure 3E**).

The class of reverse transcriptase enzyme impacts the mismatch frequency and mutational signature

The retroviral reverse transcriptase SuperScript-III (SS3) enzyme is the most commonly used RT enzyme in next-generation sequencing library preparations, including RNA-Seq and

DMS-Seq. However, in the presence of modified RNA nucleotides that affect Watson-Crick base pairing, retroviral reverse transcriptases lead to a large proportion of RT drop-off compared to mismatched nucleotide incorporation [25, 33]. In contrast, the thermostable group II intron reverse transcriptase (TGIRT) enzymes have higher processivity, fidelity, and thermostability than retroviral RTs [34], leading to lower RT drop-off rates [33], and it has been shown that TGIRT increases the mismatch frequency in DMS-modified datasets [25].

Previous studies pioneering DMS-MaPSeq employed TGIRT to increase the proportion of mismatched positions in their datasets, and identified 51% of the DMS-modified positions to be adenosines [25]. This finding is in contrast with our results, where our mutational profiling identifies 66% of DMS-probed positions at cytosines, and only 23% at adenosines (**Figure 2E**). We hypothesized that the choice of reverse transcriptase enzyme may not only increase the number of mismatches, but also unequally increase the mismatch frequency of m^1A positions compared to m^3C positions. To test this hypothesis, we compared DMS-modified 64c zebrafish embryo datasets, where samples had been reverse transcribed using either SS3 or TGIRT employing a Structure-Seq protocol. We find that the proportion of reference nucleotides in DMS-modified samples is largely affected by the enzyme used for reverse transcription (**Figure 4A**).

In agreement with this observation, we find that the increased processivity of TGIRT is not equal across DMS-modified positions; more specifically, m^1A positions exhibit higher processivity than m^3C positions (**Figure 4B**). Consequently, the mismatch signal from the TGIRT experiment was enriched at adenosines, whereas the SS3 experiment displayed a stronger signal at cytosines (**Figure 4A)**. Interestingly, the relative proportions of reference nucleotides of SS3 samples that have been analyzed using RT-stop methodologies, are highly similar to those that have been reverse transcribed using TGIRT and analyzed using MaP (**Figure 4C**). This coincidental similar proportion of relative reference nucleotides may explain why prior works had not further characterized the non-overlapping nature of the two methodologies.

We finally wondered whether the mismatch signatures revealed by our analysis was conserved when using different RT enzymes. We find that the mismatch signatures observed in Structure-Seq datasets are similar to those found in DMS-Seq datasets using SS3 for both m^1A and m^3C (**Figures 4D and 4E**). In contrast, the mismatch signature drastically changes when using the TGIRT enzyme. Specifically, m^1A favors a substitution for a thymine while m^3C doesn't exhibit such a clear preferential mutational signature as with SS3 (**Figure 4D**

and 4F). Altogether, RT enzymes alter the processivity of the enzyme unequally across RNA modifications, and also affect the proportion and identity of the nucleotides that are misincorporated. Hence, we propose that mismatch signatures are not an intrinsic property of RNA modifications, but rather, they are specific to each RNA modification-reverse transcriptase combination.

DISCUSSION

Next-generation sequencing (NGS) has revolutionized the field of molecular biology, opening new avenues to explore the genome, epigenome and transcriptome. In the last few years, genome-wide techniques to explore additional layers of regulation, such as RNA structure or the epitranscriptome, have become available. Due to their relatively recent appearance, we are still facing the challenges of determining how to best analyze these data, as well as properly interpreting the results.

Current RNA structure studies are mainly limited to probing single-stranded (ss) regions. Thus, it is essential to maximize the analysis of such datasets and to capture the whole spectrum of the provided structural information. Here we show that mutational profiling and reverse truncation signals are both valid methodologies to identify DMS-modified positions, however, the two methodologies capture only part of the structural information. Although there is a significant overlap of identified positions between the two methodologies (**Figures 1D, 1E, 2A-C and S1**), the quantitative correlation between mismatch frequencies and RT drop-offs (accessibilities) is nonexistent or even negative (**Figures 2C** and **D**). Therefore, contrary to current practice, we suggest that the optimal identification of RNA structure is generated by the union of mismatch and RT drop-off signals.

Upon DMS treatment, single-stranded RNA undergoes methylation in three of its four nucleotides, giving rise to m^1A, m^3C and m^7G, respectively. In addition to the methylations that occur upon DMS treatment, around 20 different naturally occurring RNA modifications can also affect reverse transcription, causing both reverse transcription truncation as well as increased mismatch rates at the modified position [28, 29, 33, 35]. Here we show that multiple mismatch positions in DMS-Seq datasets are present in RNA-seq (**Figure 3C**), and thus, unrelated to DMS probing. We show that coupling RNA-Seq to DMS-Seq allows to filter out these positions, increasing the signal-to-noise ratio. Unfortunately, RT truncations generated by naturally occurring RNA modifications will still be a source of error in our analyses, which could be partially alleviated by using highly processive reverse

transcriptases, such as thermostable group II intron reverse transcriptases [34], rather than the commonly used viral reverse transcriptases. The use of these enzymes can increase the number of identifiable positions using mutational profiling, especially m^1A positions (**Figure 2E**).

Compared to other genome-wide RNA structure probing reagents such as SHAPE, DMS is especially valuable to optimize the filtering step, as it only modifies A and C bases, thus we can assess the noise-to-signal ratio based on the number of predicted DMS-modified positions that fall in G and T bases. Although this work is mainly focused on the analysis of DMS-modified datasets, our findings should be applicable to any probing methodology. In a similar fashion to DMS-modified samples, SHAPE-modified datasets are currently being analyzed using either RT truncation methodologies [20, 36] or mutational profiling [23, 26], but not both.

Importantly, our findings are not only applicable to the field of RNA structure, but also to the field of RNA modifications. RNA modifications are known to modulate the structure, function and activity of RNA molecules [37-43]. Recent papers have analyzed multiple RNA-Seq datasets looking for mismatched nucleotide signatures, finding that many RNA sequences contain modified nucleosides, [28, 29, 44], and in the last year, genome-wide maps identifying thousands of m^1A modifications have been made available [45]. However, previous studies have been mainly utilizing mutational profiling as a means to identify RNA modifications from RNA-Seq datasets [28]. Therefore, for a more complete identification of genome-wide RNA modifications, we suggest that a combination of both RT truncation and mutational profiling methodologies should also be employed.

Furthermore, our analysis shows that the choice of the RT enzyme does not only affect the mismatch/RT drop-off ratio, but also affects the relative proportion of misincoporated nucleotides, dramatically affecting the mismatch signatures. Consequently, mismatch signatures obtained with a given RT enzyme (e.g. SS3) cannot be used to predict DMS-modified positions in datasets that have been reverse-transcribed with a different RT enzyme (e.g. TGIRT). This statement is also true for analyses aiming to detect naturally occurring RNA modifications based on mismatch signature [28, 29, 44]. Whether additional variables, such as RT temperature or salt concentration, may affect the mismatch signature of RNA modifications is still an open question.

Overall, here we show that the DMS signal cannot be entirely captured neither qualitatively nor quantitatively using only mutational profiling or RT drop-off methodologies. Indeed, not only DMS-modified positions are exclusively identified by one of the two methodologies (**Figures 2B-2C**), but also, amongst the positions identified by both methodologies, the level of modification does not correlate with each other (**Figure 2D**), suggesting that the actual level of modification is the sum of the two methodologies. On the other hand, we also find that mutational profiling preferentially identifies DMS-modified C bases (m^3C) whereas RT drop-off methodologies preferentially identify DMS-modified A bases (m^1A) (**Figures 2C and 2E**). Therefore, the relative proportion of mismatch/RT drop-off is dependent on the identity of the modified base, as well as on the sequence context. While more processive RT enzymes, such as TGIRT, can increase RT processivity, here we show that this increased processivity is not equal across all RNA modifications. Consequently, we propose that a combination of both MaP and RT drop-off signals should be employed to obtain the most from genome-wide RNA structure probing datasets, regardless of the reverse transcriptase employed during the library preparation.

MATERIALS AND METHODS

Zebrafish maintenance

Wild-type zebrafish embryos were obtained through natural mating of TU-AB strain of mixed ages (5-18 months). Mating pairs were randomly chosen from a pool of 70 males and 70 females allocated for each day of the month. Fish lines were maintained following the International Association for Assessment and Accreditation of Laboratory Animal Care research guidelines, and approved by the Yale University Institutional Animal Care and Use Committee (IACUC).

DMS-Seq, RNA-Seq and Structure-Seq datasets

DMS-Seq datasets of *in vitro* DMS treated known RNA structure spike-ins (*Tetrahymena* ribozyme and DsRed mRNA containing a tRNA-spinach cassette in its 3'UTR) and *in vivo* DMS treated 64c stage zebrafish embryos (samples: SRS2404542 and SRS2404544) were taken from SRP114782. RNA-Seq datasets from 64c stage zebrafish embryos (samples: SRS2404514 and SRS2404517) were also taken from SRP114782. The manuscript associated to SRP114782 is currently under review and all samples will be released immediately after acceptance. In the meantime, all datasets mentioned above related to SRP114782 are available upon request. Structure-Seq datasets of 64c stage zebrafish

embryos using either SSIII or TGIRT reverse transcription enzymes are accessible at SRP115809. See Table S1 for more details on all datasets.

Structure-Seq experiments

For *in vivo* modification of zebrafish embryo transcriptome, 150 64-cell embryos were transferred to 5 mL eppendorf tubes containing 400 µL of system water from the fish facility. 100% DMS (Sigma-Aldrich) was diluted in 100% ethanol to obtain a 20% DMS stock solution. The DMS stock solution was then diluted to 6% DMS in 600 mM Tris-HCl pH 7.4 (AmericanBio) in system water from the fish facility. This DMS/Tris-HCl solution was immediately mixed vigorously and 200 µL was added to each embryo containing tube to reach a final concentration of 2% DMS and 200 mM Tris HCl pH 7.4. Embryos were incubated at room temperature for 10 min with occasional gentle mixing. The DMS solution was then quickly removed from the tubes and the embryos were flash frozen in liquid nitrogen. Frozen embryos were thawed and actively lysed with 800 µL of TRIzol (Life Technologies) supplemented with 0.7 M ß-mercaptoethanol (Sigma-Aldrich) to quench any remaining trace of DMS. After 2 min incubation, TRIzol was added to reach a final volume of 4 mL and total RNA extracted following the manufacturer's protocol. Poly(A)+ transcripts were purified using oligo $d(T)_{25}$ magnetic beads (New England BioLabs) following the manufacturer's protocol and eluted in 35 µL of water. DMS treatments were performed in duplicate from different clutches and days. For each replicate, an untreated control was performed following the same steps, omitting the DMS the different solutions.

Structure-Seq libraries were prepared similar to Ding *et al.* [13] with few changes. Briefly, DMS treated or untreated poly(A)+ RNA duplicates were pooled together and subjected to reverse transcription using a partially degenerated primer fused with part of an Illumina TruSeq adapter (5'- AGACGTGTGCTCTTCCGATCTNNNNNN-3') and either the SuperScript III First Strand Kit (Invitrogen) or the TGIRT^{TM}-III (InGex) following manufacturer' protocols. Therefore, each type of reverse transcription reactions was performed using the same pool of DMS-modified RNAs allowing a direct comparison of the two enzymes. For reverse transcription reactions with the SuperScrit III enzyme, samples were heated at 25°C for 10 min, 42°C for 30 min, 50°C for 10 min, 55°C for 20 min, and 75°C for 15 min to deactivate the enzyme. For reverse transcription reactions with the TGIRT^{TM}-III enzyme, samples were heated at 25°C for 10 min, 42°C for 10 min, 50°C for 10 min, 55°C for 10 min, 60°C for 30 min, 65°C for 20 min, and 75°C for 15 min to deactivate the enzyme. All samples were treated with RNAse H at 37°C for 20 min. cDNAs were purified using 36 µL of Agencourt

AMPure XP beads (Beckman Coulter) following manufacturer' protocol and resuspended in 10 µL of water. ssDNA linker (/5Phos/NNNNNGATCGTCGGACTGTAGAACTCTGAAC/3InvdT/) was ligated at cDNA 3'-ends using the CircLigase ssDNA ligase (Epicentre) with slightly modifications to the manufacturer's protocol, i.e. where the different reagents were added to 3 µL of cDNAs to reach the following final concentrations: 50 units of CircLigase, 0.05 mM ATP, 2.5 mM $MnCl_2$, 10% PEG 6000, 1 M betaine, and 5 µM ssDNA linker in a final volume of 10 µL. Ligation reactions were incubated at 60°C for 2h, 68°C for 1h, and 80°C for 10 min to deactivate the ligase. 10 µL of water was added to each reaction. The resulting 20 µL ligation products were purified using 36 µL of Agencourt AMPure XP beads and dissolved in 16 µL of water. PCR amplification was performed on the ligated cDNA using Illumina primers (Small RNA PCR Primer 2 5'-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3' and PCR primer index 5'-CAAGCAGAAGACGGCATACGAGATbarcodeGTGACTGGAGTTCAGACGTGTGCTCTTCC GATCT-3', where barcode is the 6-nucleotide index). PCR products were purified and concentrated using MinElute PCR Purification Kit (QIAGEN) and eluted in 10 µL of water. Eluted PCR products were separated in a 2% agarose gel and products of 200-1,000 nucleotide were extracted and purified using the MinElute Gel Extraction Kit (QIAGEN). Libraries were sequenced on Illumina HiSeq 2000/2500 machines producing single-end 76 nucleotide reads. Sequencing samples are summarized in Table S1.

Read filtering and mapping

DMS-seq raw reads contained the following features: NNNN-insert-NN-barcode(4-mer)-adapter where the 6N (NNNN+NN) sequence composes the Unique Molecular Identifier (UMI), "barcode" is the sample 4-mer in-house barcode and adapter is the 3'-illumina adapter. The UMI was used to discard PCR duplicates and count single ligation event. The barcode was used to mark individual replicate following the 3'-adapter ligation step. Base calling was performed using CASAVA-1.8.2. The Illumina TruSeq index adapter sequence was then trimmed by aligning its sequence, requiring 100% match of the first five base pairs and a minimum global alignment score of 60 (Matches: 5, Mismatches: -4, Gap opening: -7, Gap extension: -7, Cost-free ends gaps). Trimmed reads were demultiplexed based on the sample's in-house barcode, the UMI was clipped from the 5'- and 3'-end and kept within the read name, for marking PCR duplicates. Structure-Seq reads contained the following features: NNNNN-insert, where 5N correspond to the UMI, and were processed as for the DMS-Seq reads omitting the Illumina index adapter trimming and clipping only the 5'-end UMI. DMS-Seq and Structure-Seq reads were then depleted of rRNA, tRNA, snRNA,

snoRNA and miscRNA, using Ensembl78 annotations, as well as from RepeatMasker annotations, using strand-specific alignment with Bowtie2 v2.2.4 [46]. The remaining reads were aligned to the zebrafish Zv9 genome assembly using STAR version 2.4.2a [47] with the following non-default parameters: *--alignEndsType EndToEnd --outFilterMultimapNmax 100 --seedSearchStartLmax 15 --sfbdScore 10 --outSAMattributes All*. Genomic sequence indices for STAR were built including exon-junction coordinates from Ensembl 78. Only reads of unique UMI were kept at each genomic coordinate for DMS-seq and ribosome profiling experiments. Raw reads from RNA-seq experiments were processed using the same pipeline, omitting the adapter trimming, barcoding demultiplexing and UMI clipping steps. The filtered reads were aligned onto Zebrafish Zv9 assembly using STAR, with the same parameters as described above. STAR genomic sequence indices were built including exon-junction coordinates from Ensembl 78.

Analysis of accessibility (RT-stop methodology)

Per-transcript profiles were computed using uniquely mapped reads overlapping at least 10 nucleotides with the transcripts annotation. Each read count was attributed to the nucleotide in position –1 of the read's 5'-end within the transcript coordinate, to correct for the fact that reverse transcription stops one nucleotide prior to the DMS-modified nucleotide. To determine read distributions for each nucleotide, only transcripts with a minimum of 100 counts were considered. Accessibilities were calculated following the 2%-8% rule [48], i.e. by normalizing the read counts proportionally to the most reactive As and Cs within the region after the removal of outliers. More specifically, the 2% most reactive As and Cs were discarded and each position was divided by the average of the next 8% most reactive As and Cs. Accessibilities greater than 1 were set to 1, and accessibilities for G and T were set to 0.

Analysis of mutational profiles (MaP methodology)

DMS-Seq, Structure-Seq and RNA-Seq mapped bam files were processed using in-house scripts to produce a bed file of mismatched positions, including metadata information regarding reference nucleotide, coverage (number of reads at that given base pair), and relative nucleotide frequencies at each given position (measured as total number of A, C, G, and T nucleotides, normalized by the coverage). Mismatched positions were then filtered to remove SNPs, naturally occurring RNA modifications and sequencing artifacts. Due to the low mismatch frequencies that m1A and m3C modifications cause, a minimum of reads/bp was required. Overall, the set of filtered mismatch positions met all the following criteria: i) mismatches are found in both DMS-Seq replicates; ii) they are not found in the set of replicable RNA-Seq mismatches; iii) minimum coverage of 50 reads/bp; iv) minimum of 2

mismatched reads/bp; and v) maximum of 0.25 mismatch frequency. For Structure-Seq datasets, the set of mismatched positions was filtered by the control (untreated) Structure-Seq mismatches, instead of the RNA-Seq mismatches. Accessibility values from mismatch frequencies (**Figures S1C and S1D**) were calculated using the 2%-8% rule as for the RT-stop signal.

ACKNOWLEDGEMENTS

We thank all members of the Giraldez, Mattick and Kellis labs for their valuable comments and suggestions. This research was supported by the Human Frontier Science Program (LT000307/2013-L to EMN), the Australian Research Council (DE170100506 to EMN), the Fonds de Recherche du Québec - Santé (postdoctoral fellowship to JDB), the National Health and Medical Research Council (Project Grant APP1070631 to JSM) and the National Institute of Health (grants R01 HD074078, GM103789, GM102251, GM101108 and GM081602), Pew Scholars Program in the Biomedical Sciences, March of Dimes 1-FY12-230, the Yale Scholars Program and Whitman fellowship funds provided by E. E. Just, Lucy B. Lemann, Evelyn and Melvin Spiegel, The H. Keffer Hartline and Edward F. MacNichol, Jr. of the Marine Biological Laboratory in Woods Hole, MA to AJG.

FIGURE LEGENDS

Figure 1. Mutational profiling and RT drop-off analyses contain non-overlapping information. (**A**) DMS methylates nitrogen atoms (red), some of which are involved in Watson-Crick base pairing (m^1A and m^3C), causing RT drop-off and increased mutational rates during reverse transcription. (**B**) Overview of the two major methodologies relying either on RT drop-off or mutational profiling to quantify DMS signal. (**C**) Replicability of mutational profiles and RT drop- off accessibilities of two known RNA structures. (**D and E**) Correlation between mutational profiles (“mismatch”, black) and RT stops (“accessibility”, blue) in the *Tetrahymena* ribozyme (D) and the DsRed mRNA containing a spinach tRNA cassette in its 3’UTR (E) spike-ins. Overlapping positions (“common”) that are detected by both methodologies are depicted in orange.

Figure 2. DMS-mediated mutational profiling gives reliable structural information and is influenced by the nature of the DMS-modified nucleotide. (**A and B**) Overlap of DMS-modified positions identified using mutational profiling or RT truncation methodologies in the zebrafish transcriptome (A), and the *Tetrahymena* ribozyme (B, top) and DsRed mRNA with

a spinach tRNA cassette (B, bottom) spike-ins. (**C**) Comparison of the mismatch frequencies (MaP) and the accessibilities (RT-stop) onto the experimentally determined RNA secondary structure of the domain P4-5-6 of the *Tetrahymena* ribozyme. Nucleotide positions that are preferentially identified by MaP are coloured in red, while those that are preferentially identified by RT stops are coloured in green. Those equally identified by both approaches are coloured in yellow. See also Figure S1. (**D**) Correlation between the accessibility and the mismatch frequency of the DMS-modified positions detected ("common") by both methodologies in the two spike-ins. Each nucleotide is coloured according to its reference nucleotide. The area of the plot is divided into 3 regions: *i*) mismatch preference (high mismatch frequency, low accessibility), shaded in red; *ii*) RT-stop preference (high accessibility, low mismatch frequency), shaded in green; and *iii*) no preference, shaded in orange. (**E**) Proportion of reference nucleotides in transcriptome-wide DMS-seq experiments from 64c stage zebrafish embryos, using the RT-stop (left) or the mutational profiling (right) methodologies**.** (**F**) Comparison of the sequence context of mismatched positions and RT-stop positions in DMS-probed 64c-stage zebrafish embryos.

Figure 3. Mismatch signature analysis of DMS signal across a vertebrate transcriptome. (**A and B**) Replicability of mismatch frequencies for both DMS-Seq (A) and RNA-Seq (B) datasets, using raw mismatches (no filtering, left panels) and filtered mismatches (right panels, see Methods). (**C**) Mismatch positions (frequency > 0.01) along exon 9 of the RPL4 gene, in DMS-Seq and RNA-Seq datasets. Mismatches identified by each individual replicate are shown. Highlighted in green squares are those positions that have mismatch frequency greater than 1% in both RNAseq and DMSSeq, and that are consistent across replicates. (**D and E**) Ternary plots highlighting the mismatch signature found in DMS-modified transcriptome. For each reference nucleotide, the relative proportion of the other three nucleotides at mismatched positions is shown. The top (D) and bottom (E) rows depict transcriptome-wide DMS-Seq mismatches found in zebrafish 64c-stage embryos, pre- and post-filtering. The right column shows the frequency of reference nucleotides at mismatch positions.

Figure 4. **Mutational profiling signatures differ between reverse transcriptases.** (**A**) Proportion of reference nucleotides in transcriptome-wide Structure-seq experiments, using either the SuperScript-III (left) or TGIRT (right) reverse transcriptases. The set of DMS-modified positions has been identified using mutational profiling in both conditions. (**B**) Proportion of mismatched positions identified with the SS3 and the TGIRT reverse transcriptases. (**C**) Comparison of the relative proportion of reference nucleotides, when

using different reverse transcriptases as well as different analysis methodologies. (**D**) Proportion of misincorporated nucleotides at mismatched positions, using either the SS3 or TGIRT reverse transcriptase, at m^1A (left) and m^3C positions (right). Boxplot outliers have been removed for enhanced visualization. (**E and F**) Ternary plots showing the transcriptome-wide mismatch signatures induced by m^1A (left) and m^3C (right), when using either the SS3 (E) or the TGIRT (F) enzymes in Structure-Seq datasets.

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. (**A and B**) Comparison of the overlay of accessibilities (left) and the overlap of mismatch frequencies (right) onto two different regions of the experimentally determined RNA secondary structure of the *Tetrahymena* ribozyme. Each nucleotide has been colored based on either their normalized accessibility (left) or their mismatch frequency (right), respectively. Arrows point to examples where the nucleotide is: *i*) preferentially predicted by the RT truncation method (green), *ii*) preferentially predicted by mismatch profiling (red), or *iii*) equally predicted by both methods (orange) (**C and D**) Box plots of the agreement between accessibilities calculated from either RT stop or mismatch signals and A/C pairing statuses (ss, single-stranded; ds, double-stranded) for the *Tetrahymena* ribozyme (C) and the tRNA-Spinach cassette (D).

Figure S2. (**A and B**) Mismatch signature ternary plots for the two spike-ins, the *Tetrahymena* ribozyme and Spinach tRNA cassette, prior to filtering (A) and after filtering (B)**.** For filtering details, see Methods. (**C**) Relative proportion of misincorporated nucleotides at mismatched positions, when the reference nucleotide is A (top) or C (bottom) in the spike-ins. Boxplot outliers have been removed for enhanced visualization (**D**) Relative proportion of misincorporated nucleotides at mismatched positions, when the reference nucleotide is A (top) or C (bottom) in transcriptome-wide DMS-probed 64c stage zebrafish embryos. Outliers have been removed for better visualization of the boxplot differences.

Figure S3. Mismatch signature plots of 64c zebrafish transcriptome-wide DMS-Seq samples, for each reference nucleotide. The relative proportion of the other three nucleotides at mismatched positions is shown as a ternary plot. Ternary plots have been computed at each stage of the filtering that has been performed. The proportion of reference nucleotides after each filtering step is depicted on the right side of the figure.

REFERENCES

1. Mortimer, S.A., M.A. Kidwell, and J.A. Doudna, *Insights into RNA structure and function from genome-wide studies.* Nat Rev Genet, 2014. **15**(7): p. 469-79.
2. Sharp, P.A., *The centrality of RNA.* Cell, 2009. **136**(4): p. 577-80.
3. Morris, K.V. and J.S. Mattick, *The rise of regulatory RNA.* Nat Rev Genet, 2014. **15**(6): p. 423-37.
4. Mercer, T.R. and J.S. Mattick, *Structure and function of long noncoding RNAs in epigenetic regulation.* Nat Struct Mol Biol, 2013. **20**(3): p. 300-7.
5. Dethoff, E.A., et al., *Functional complexity and regulation through RNA dynamics.* Nature, 2012. **482**(7385): p. 322-30.
6. Geisberg, J.V., et al., *Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast.* Cell, 2014. **156**(4): p. 812-24.
7. Kertesz, M., et al., *Genome-wide measurement of RNA secondary structure in yeast.* Nature, 2010. **467**(7311): p. 103-7.
8. Wan, Y., et al., *Landscape and variation of RNA secondary structure across the human transcriptome.* Nature, 2014. **505**(7485): p. 706-9.
9. Tijerina, P., S. Mohr, and R. Russell, *DMS footprinting of structured RNAs and RNA-protein complexes.* Nat Protoc, 2007. **2**(10): p. 2608-23.
10. Ehresmann, C., et al., *Probing the structure of RNAs in solution.* Nucleic Acids Res, 1987. **15**(22): p. 9109-28.
11. Weeks, K.M., *Advances in RNA structure analysis by chemical probing.* Curr Opin Struct Biol, 2010. **20**(3): p. 295-304.
12. Bevilacqua, P.C., et al., *Genome-Wide Analysis of RNA Secondary Structure*, in *Annu. Rev. Genet.* 2015. p. annurev-genet-120215-035034.
13. Ding, Y., et al., *Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq.* Nat Protoc, 2015. **10**(7): p. 1050-66.
14. Kwok, C.K., et al., *The RNA structurome: transcriptome-wide structure probing with next-generation sequencing*, in *Trends in Biochemical Sciences*. 2015, Elsevier Ltd. p. 221-232.
15. Rouskin, S., et al., *Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo.* Nature, 2014. **505**(7485): p. 701-5.
16. Watters, K.E., et al., *Characterizing RNA structures in vitro and in vivo with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq).* Methods, 2016. **103**: p. 34-48.
17. Chan, D., C. Feng, and R.C. Spitale, *Measuring RNA structure transcriptome-wide with icSHAPE.* Methods, 2017.
18. Flynn, R.A., et al., *Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE.* Nat Protoc, 2016. **11**(2): p. 273-90.
19. Fang, R., et al., *Probing Xist RNA Structure in Cells Using Targeted Structure-Seq.* PLoS Genet, 2015. **11**(12): p. e1005668.
20. Mortimer, S.A., et al., *SHAPE-Seq: High-Throughput RNA Structure Analysis.* Curr Protoc Chem Biol, 2012. **4**(4): p. 275-97.
21. Chen, D. and J.T. Patton, *Reverse transcriptase adds nontemplated nucleotides to cDNAs during 5'-RACE and primer extension.* Biotechniques, 2001. **30**(3): p. 574-80, 582.
22. Fuchs, R.T., et al., *Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure.* PLoS One, 2015. **10**(5): p. e0126049.
23. Siegfried, N.A., et al., *RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP).* Nat Methods, 2014. **11**(9): p. 959-65.
24. Smola, M.J., et al., *Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis.* Nat Protoc, 2015. **10**(11): p. 1643-69.

25. Zubradt, M., et al., *DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo.* Nat Methods, 2017. **14**(1): p. 75-82.
26. Smola, M.J., et al., *SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells.* Proc Natl Acad Sci U S A, 2016. **113**(37): p. 10322-7.
27. Li, X., et al., *Transcriptome-wide mapping reveals reversible and dynamic N(1)-methyladenosine methylome.* Nat Chem Biol, 2016. **12**(5): p. 311-6.
28. Ryvkin, P., et al., *HAMR: high-throughput annotation of modified ribonucleotides.* RNA, 2013. **19**(12): p. 1684-92.
29. Hauenschild, R., et al., *The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent.* Nucleic Acids Res, 2015. **43**(20): p. 9950-64.
30. Golden, B.L., et al., *A preorganized active site in the crystal structure of the Tetrahymena ribozyme.* Science, 1998. **282**(5387): p. 259-64.
31. Warner, K.D., et al., *Structural basis for activity of highly efficient RNA mimics of green fluorescent protein.* Nat Struct Mol Biol, 2014. **21**(8): p. 658-63.
32. Aird, D., et al., *Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.* Genome Biol, 2011. **12**(2): p. R18.
33. Zheng, G., et al., *Efficient and quantitative high-throughput tRNA sequencing.* Nat Methods, 2015. **12**(9): p. 835-7.
34. Mohr, S., et al., *Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing.* RNA, 2013. **19**(7): p. 958-70.
35. Dai, Q., et al., *Selective Enzymatic Demethylation of N2 ,N2 -Dimethylguanosine in RNA and Its Application in High-Throughput tRNA Sequencing.* Angew Chem Int Ed Engl, 2017. **56**(18): p. 5017-5020.
36. Incarnato, D., et al., *RNA structure framework: automated transcriptome-wide reconstruction of RNA secondary structures from high-throughput structure probing data.* Bioinformatics, 2016. **32**(3): p. 459-61.
37. Novoa, E.M., C.E. Mason, and J.S. Mattick, *Charting the unknown epitranscriptome.* Nat Rev Mol Cell Biol, 2017. **18**(6): p. 339-340.
38. Nainar, S., et al., *Evolving insights into RNA modifications and their functional diversity in the brain.* Nat Neurosci, 2016. **19**(10): p. 1292-8.
39. Agris, P.F., *The importance of being modified: an unrealized code to RNA structure and function.* RNA, 2015. **21**(4): p. 552-4.
40. Gilbert, W.V., T.A. Bell, and C. Schaening, *Messenger RNA modifications: Form, distribution, and function.* Science, 2016. **352**(6292): p. 1408-12.
41. Satterlee, J.S., et al., *Novel RNA modifications in the nervous system: form and function.* J Neurosci, 2014. **34**(46): p. 15170-7.
42. Zhang, X., et al., *Small RNA Modifications: Integral to Function and Disease.* Trends Mol Med, 2016. **22**(12): p. 1025-1034.
43. Huisman, B., et al., *Functional Dissection of the m6A RNA Modification.* Trends Biochem Sci, 2017. **42**(2): p. 85-86.
44. Vandivier, L.E., et al., *Chemical Modifications Mark Alternatively Spliced and Uncapped Messenger RNAs in Arabidopsis.* Plant Cell, 2015. **27**(11): p. 3024-37.
45. Dominissini, D., et al., *The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA.* Nature, 2016. **530**(7591): p. 441-6.
46. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
47. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
48. Deigan, K.E., et al., *Accurate SHAPE-directed RNA structure determination.* Proc Natl Acad Sci U S A, 2009. **106**(1): p. 97-102.

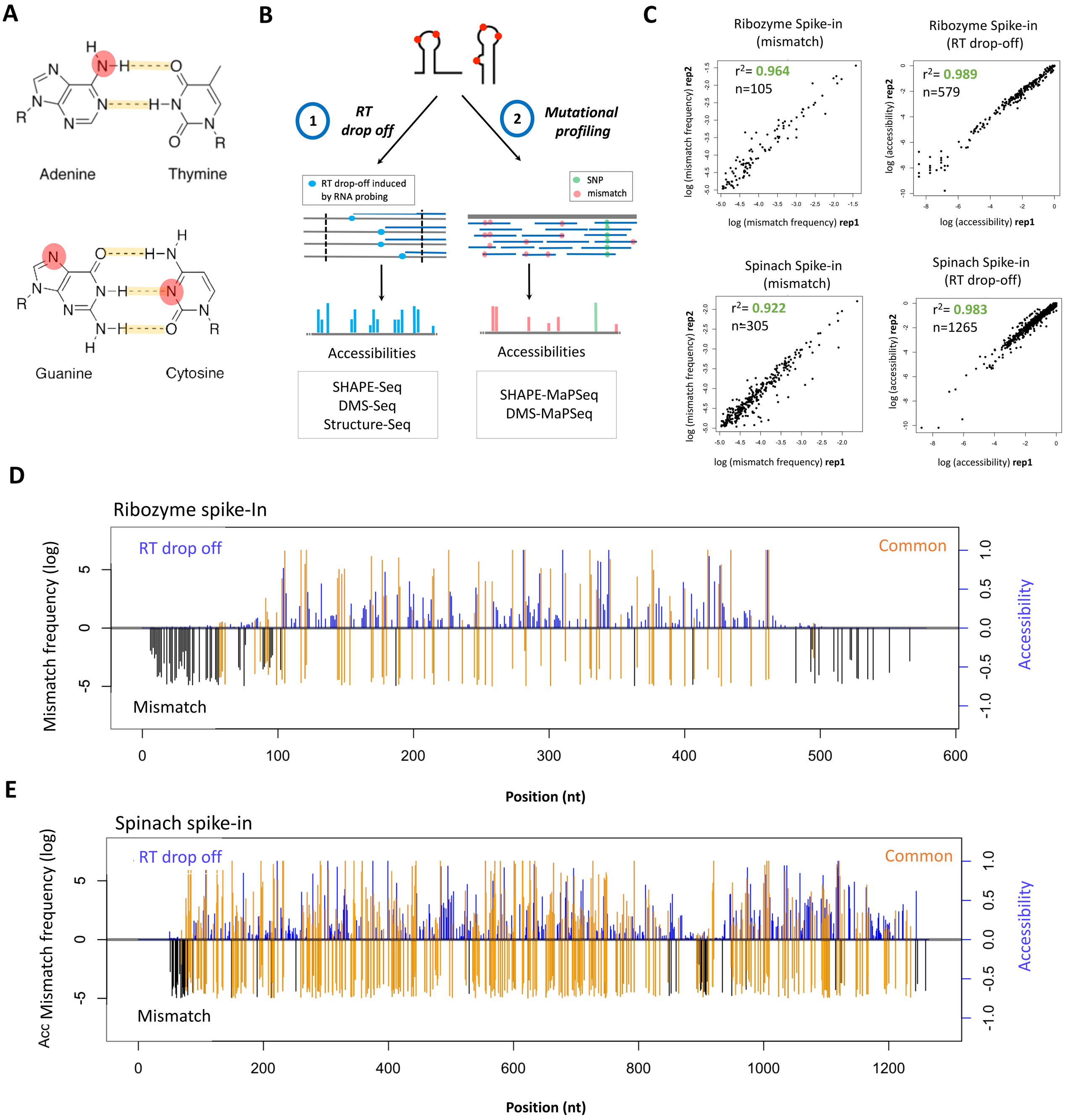

Figure 1. Novoa et al. 2017

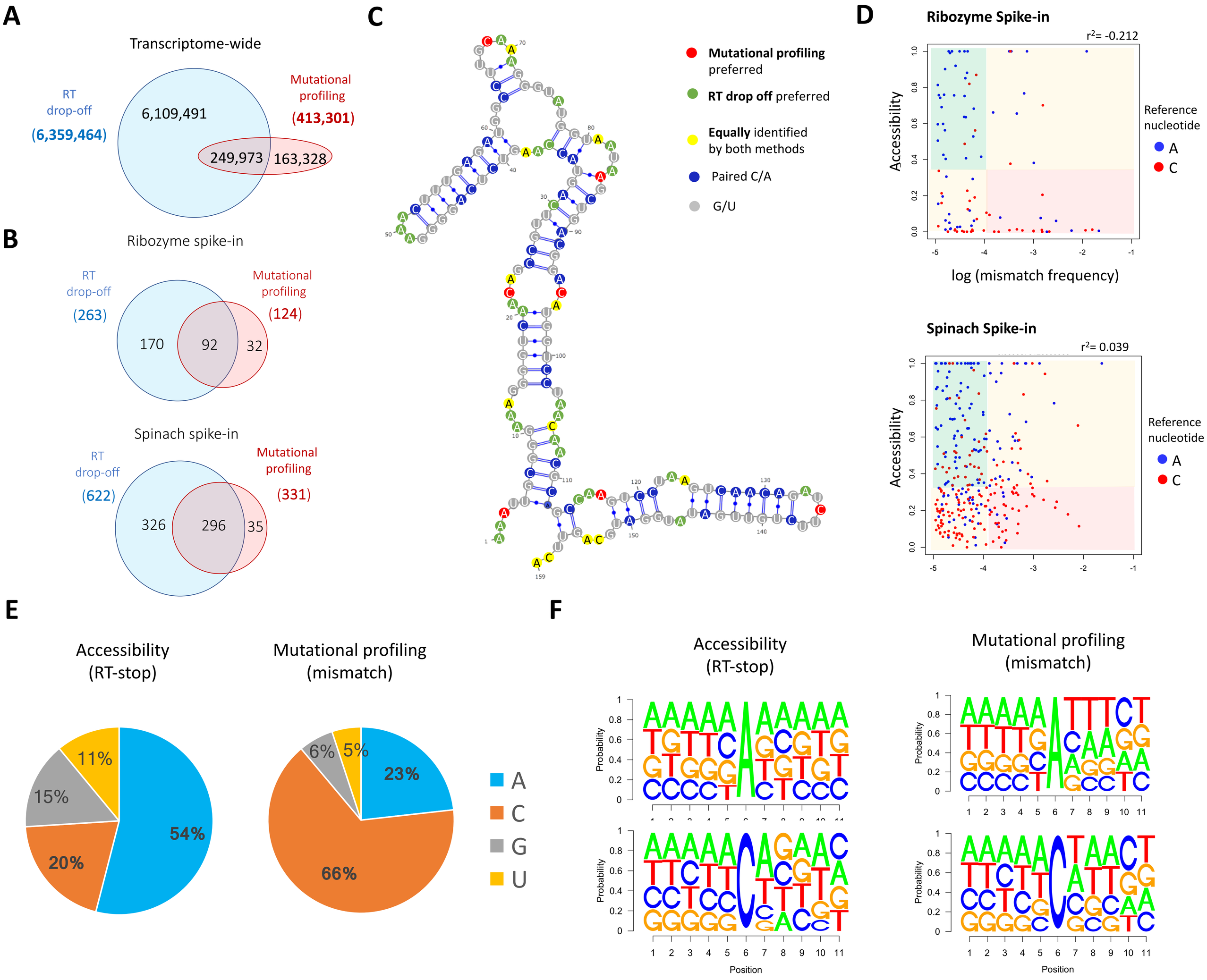

Figure 2. Novoa et al. 2017

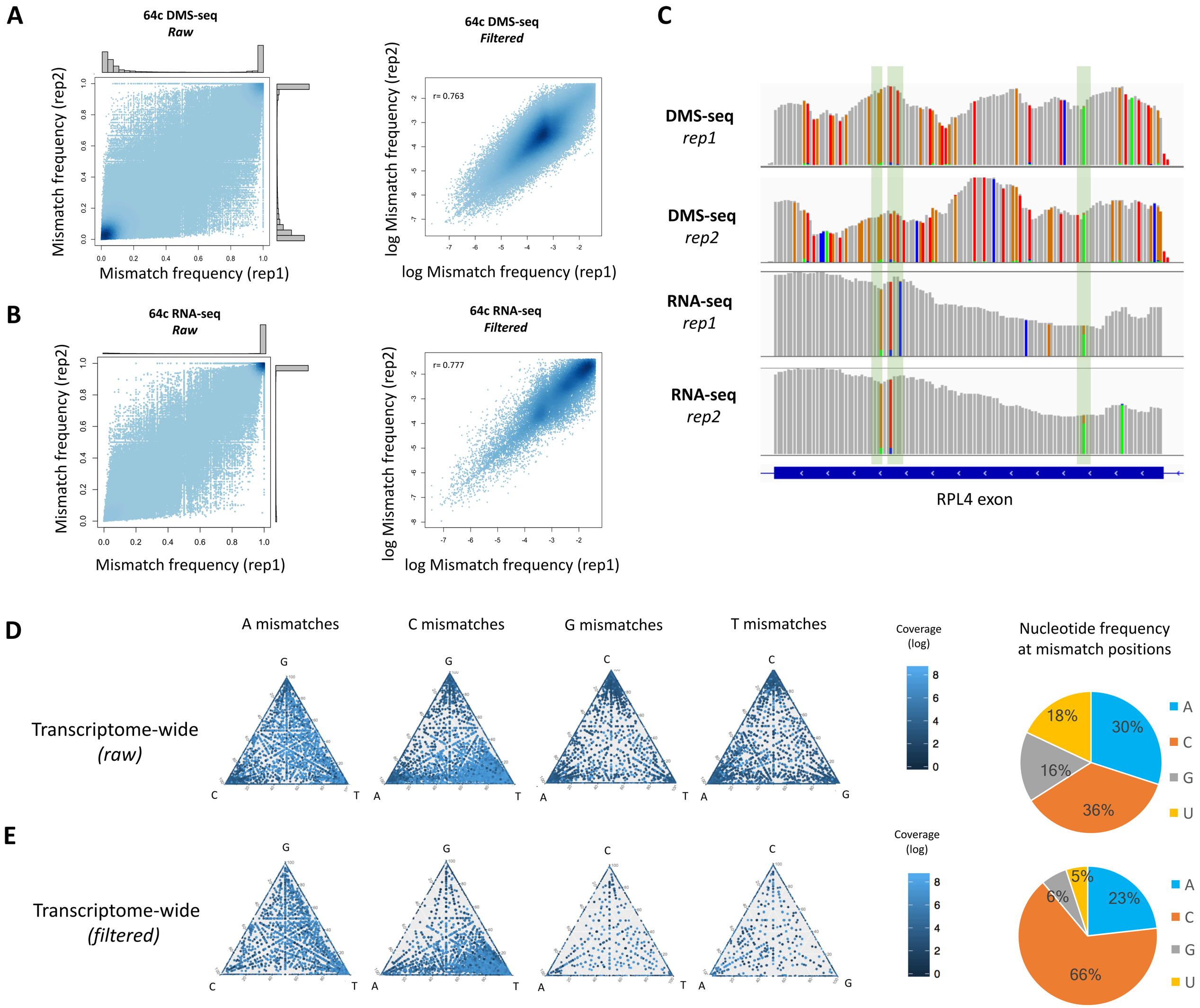

Figure 3. Novoa et al. 2017

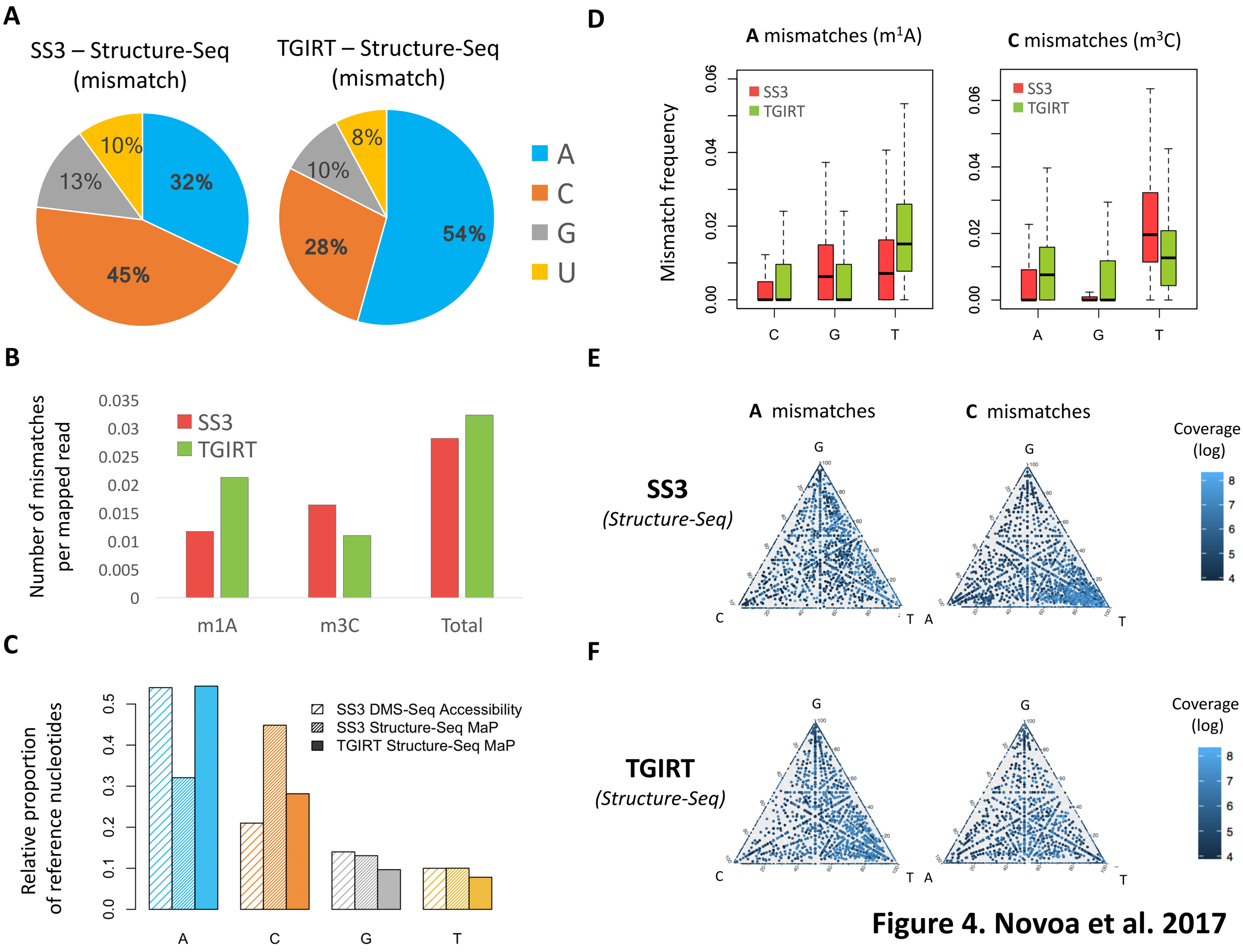

Figure 4. Novoa et al. 2017