Title

Novel methods for the quantitative determination of RNA folding on a genome-wide scale and in a targeted manner

Author

Zubradt, Meghan

Publication Date

2016

Novel methods for the quantitative determination of RNA folding on a genome-wide scale and in a targeted manner

by

Meghan McKeon Zubradt

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

ACKNOWLEDGEMENTS

The mentoring and support I received throughout this thesis project were paramount to its success. Of particular note is Silvia Rouskin, who trained, mentored, and encouraged me while we overlapped in the lab and continued to support me as a collaborator afterwards. Her creative ideas and optimism were the basis for much of our success.

Jonathan Weissman created an incredible environment for scientific investigation, supported my ideas, provide invaluable advice and insights, and ensured the lab was an enjoyable place to be everyday with generous, fun, intelligent, and helpful colleagues.

Finally, I have to thank my friends and family—of which there are too many to mention each by name—who provided balance and perspective needed during the more challenging times. Specifically, to my husband Brian, thank you for ensuring that my work life was always balanced by a real life that was fun, exciting, interesting, and a joy to spend with you.

Novel methods for the quantitative determination of RNA folding on a genome-wide scale and in a targeted manner

Meghan McKeon Zubradt

ABSTRACT

RNA has many important functions in the cell beyond its role as a molecular intermediate between DNA and proteins. These functions include scaffolding, catalysis, localization, translation regulation, and more. It is the ability of RNA to base pair and form secondary and complex tertiary structures that is the source of its biological versatility, but until recently, the research community lacked tools to study *in vivo* RNA folding in a high-throughput or quantitative manner. The recent coupling of chemical RNA structure probing with a next-generation sequencing (NGS) readout has enabled great strides forward in our understanding of cellular folding dynamics and the *de novo* discovery of RNA secondary structures. However, these first generation approaches have limitations due to the readout of chemical modifications as truncation products generated during reverse transcription. These limitations include cryptic biases introduced during library generation and a reliance on RNA structure signal derived from a population average. This latter limitation prevents the investigation of *in vivo* RNA structure heterogeneity within an RNA species, which is a fascinating yet entirely open question in the field. To resolve the limitations of previous techniques and to broadly enable the *in vivo* investigation of RNA structure diversity, we developed a second-generation NGS-coupled RNA structure probing approach called dimethyl sulfate mutational profiling with sequencing (DMS-MaPseq). DMS-MaPseq is based on encoding RNA structure-specific dimethyl sulfate modifications as mismatches during reverse transcription, instead of as truncation products. This technical change yields excellent data that does not require correction or comparison to a background control sample, and, importantly, is compatible with a genome-wide or target-specific amplification. Importantly, target-specific amplification allows for the investigation of low-expression RNAs that do not receive sufficient coverage in genome-wide samples, and this strategy has allowed us to identify a novel structural element in the human FXR2 5′ UTR that regulates translation in *cis*. Our targeted strategy is also more cost-effective and technically accessible than its genome-wide counterpart, but we also demonstrate its much broader utility in the investigation of RNA structure heterogeneity. Specifically, we use DMS-MaPseq to investigate the RNA structure variation in human alleles and to assess RNA structure differences *in vivo* in pre-mRNA versus its mature processed counterpart. Finally, DMS-MaPseq provides an essential technical foundation for single-molecule RNA structure determination because it can encode multiple pieces of structural information per cDNA fragment prepared and sequenced. In summary, DMS-MaPseq enables the collection of high-quality RNA structure data, allows for experimentation on a genome-wide or target-specific scale, and provides the critical framework for the investigation of RNA structure heterogeneity *in vivo*.

CONTRIBUTIONS

Excluding the contributions specified here, the work presented in this dissertation constitutes the work of Meghan McKeon Zubradt. Portions of Chapter 1 is reproduced from a publication in Nature in 2014 entitled “Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*,” and portions of Chapters 2-5 are reproduced from a publication under revision at Nature Methods entitled “DMS-MaPseq: A genome-wide or targeted approach for RNA structure probing *in vivo*.” As this latter publication is under review, portions of the text here may differ from the final published version. Paromita Gupta conducted many target-specific RNA structure experiments, including many of those presented in Chapters 3 and 4 that were prepared with our Tagmentation strategy. Sitara Persad also assisted in our genome-wide data analyses from HEK 293T cells, producing Figures 2-3b and 3-1b. All of the work presented here was completed in close intellectual and experimental collaboration with Silvia Rouskin and was supervised by Jonathan S. Weissman.

TABLE OF CONTENTS

LIST OF FIGURES & TABLES

CHAPTER ONE

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

CHAPTER ONE

The need for high-throughput RNA structure probing and the current methods available

INTRODUCTION

RNA is a functionally diverse molecule that can carry genetic information and utilize base-pairing interactions to enact many additional and varied biological processes. Over the past two decades, this broad role of RNA in the cell has become increasingly appreciated, through discoveries of RNA interference[1,2] and the identification of non-coding RNAs with important implications in human disease[3]. Mediated by intra- and intermolecular base-pairing interactions, RNA can fold into complex secondary and tertiary structures that provide the basis for much of its non-protein coding cellular activities. Examples of structure-mediated functions are numerous yet likely represent only a small fraction of functional RNA structures that exist. Included is these structure-function examples are: mechanisms of translational control, such as the yeast *HAC1* gene and global observations of regulation in *E. coli* [4,5]; post-translational modifications, such as mammalian selenocysteine insertion elements[6]; mRNA localization elements, like those in the yeast *ASH1* and Drosophila melanogaster *oskar* 3′ UTRs[7–9]; catalysis, in the case of the ribosomal RNA and metabolite-controlled riboswitches in bacteria[10]; and scaffolding functions enacted by lncRNAs[11]. These functional RNA structure examples have traditionally been identified through in-depth mechanistic investigation of an RNA of interest, which prompts the question: How many more functional RNA structures exist that we haven't identified yet?

A more complete annotation and discovery of functional RNA structures depends on the availability of high-throughput, accurate, and accessible RNA structure determination methods, particularly in an *in vivo* setting. Sequence information alone is generally not sufficient to predict RNA structure, but in combination with experimental structure data, an accurate assessment of RNA folding status can often be obtained and novel RNA structures discovered. RNA structure information can be obtained by three main classes of experimental methods: 1) the structure-specific chemical modification of RNA structure[12–17], 2) structure-specific RNase digestion[18–23], and 3) partial digestion / ligation approaches to identify proximal RNA strands[24–27]. All of these methods have recently been adapted for high-throughput and transcriptome-wide use with next generation sequencing, yet each has its advantages and drawbacks. Partial digestion / ligation approaches are exceptionally useful in the identification of novel long-range and intermolecular RNA-RNA interactions, for example, but have low resolution regarding the boundaries of the RNA structures identified. RNase digestion approaches use enzymes with excellent structure specificity for the cleavage of dsRNA species, but it can only provide *in vitro* RNA structure information, can introduce substantial biases due to cryptic cleavage site preferences, and also produces low-resolution data. We have focused our technology development efforts on the chemical modification of RNA structure, since the RNA can be effectively modified inside the cell and the data produced has single-nucleotide resolution, which greatly inform the prediction of functional RNA folds.

Dimethyl sulfate (DMS) and selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) variants have emerged as the preeminent choices for the chemical probing of RNA structure. DMS rapidly and specifically modifies unpaired adenine and cytosines *in vivo* at their Watson-Crick base-pairing positions (N1 and N3, respectively)[28], whereas SHAPE chemicals modify the 2′-OH of all four RNA nucleotides in a structure-dependent manner[29,30]. Recently, a chemical variant of the SHAPE reagent with enhanced activity *in vivo* was developed[31]. These chemical lesions are detected during cDNA synthesis as the reverse transcriptase (RT) enzyme terminates synthesis upon reaching a chemically modified nucleotide, leading to truncated cDNAs. Therefore, the site of modification is revealed by sequencing the terminal 5′ cDNA end. We and others have coupled the chemical probing of RNA structure to next-generation sequencing, allowing for the experimental analysis of RNA structure on a global scale either *in vitro* or *in vivo*[12–14,17,32]. Using our DMS-seq method *in vivo*, we revealed substantial differences in RNA structure *in vivo* versus *in vitro*, underscoring the importance of examining RNA structure in its native cellular environment[13].

In addition to important global structure observations that can be attained by genome-wide RNA structure probing methods, we used DMS-seq to identify novel and functional untranslated (UTR) structures in the yeast transcriptome[13], demonstrating the utility of genome-wide techniques in empirical RNA structure identification. To identify novel structures, we applied two statistical metrics across approximately 100 nucleotide regions of the yeast transcriptome—a Pearson's *r* value to assess the pattern variation of an *in vivo* RNA structure sample compared to a denatured control and a Gini Index, which captures unevenness in the structure signal distribution (Fig. 1-1). After identifying UTR regions in the yeast transcriptome with a low *r* value but high Gini Index, we used single-nucleotide DMS-seq data to impose constraints on the RNAfold prediction algorithm, resulting in RNA structure predictions that closely recapitulated the *in vivo* RNA structure signal. We then experimentally determined the role of these structures in controlling protein expression.

First, we identified RNA structures in the yeast *PMA1* and *SFT2* 5′ UTRs that were of particular interest given evidence for evolutionary compensatory mutations in the fungi lineage (Fig. 1-2c) and the close proximity of the structure to the AUG codon, respectively. We cloned the full 5′ UTRs upstream of a Venus fluorescent reporter (Fig. 1-2a), and then mutated the RNA structure with site-directed mutagenesis. Comparing reporter protein levels by flow cytometry for the wildtype, mutated, and then compensated versions of the *PMA1* structure revealed its role as a positive regulatory element on protein expression (Fig. 1-2b and Table1-S1). The reduction in protein levels was more drastic with an increasingly severe mutagenesis of the structure, and, interestingly, the effect on protein levels was also enhanced when yeast were grown in the cold at 10°C which has drastic thermodynamic consequences on RNA structure stability. In contrast to the *PMA1* 5′ UTR structure, mutating the *SFT2* 5′ UTR structure increased levels of the reporter protein (Fig. 1-2d,e). Restoration of the *SFT2* structure with compensatory mutations revealed a dependence on the structure stability, such that a highly stable structure proved strongly inhibitory whereas the endogenous structure had a smaller effect on the repression of protein levels.

We also investigated the phenotypic consequences of mutating a structure in the *PRC1* 3′ UTR on reporter protein levels, cloning the UTR downstream of our Venus reporter and mutating the structure in that context (Fig. 1-3a). Mutating the *PRC1* structure resulted in a drastic drop in protein levels, suggesting a role as some kind of mRNA stabilization element, and while compensatory mutations partially restored the phenotype, it did not do so fully (Fig. 1-3b,c and Table 1-S1). As a control experiment to validate our ability to identify structures from DMS-seq data, we also mutated three 3′ UTRs that did not present with strong evidence for structure *in vivo*, demonstrating no effect on protein levels when non-structured regions were perturbed (Fig. 1-3d). Our ability to identify three novel and functional RNA structures based on the empirical analysis of *in vivo* DMS-seq data demonstrates substantial progress in the *de novo* identification of biologically relevant structures from chemical-based genome-wide RNA structure probing data.

Despite important contributions to RNA structure discovery, chemical probing approaches (using either DMS or SHAPE) that rely on reverse transcriptase truncation have intrinsic limitations that render them unsuitable to address many important biological questions. For example, the heterogeneity of RNA structures *in vivo* is an important yet open question, unanswerable by current chemical probing techniques. Specifically, because only a single site of chemical modification can be observed per RNA molecule, the inferred structure corresponds to a population average that obscures any correlated signal variation on single RNA molecules.

Additionally, inherent data biases reduce the value of these approaches, especially for highly quantitative applications such as structure prediction algorithms that utilize experimental data. Included in these biases are the signal degradation that occurs when modifications are proximal to each other, as well as known enzymatic biases that can alter the capture efficiency of the information-encoding 5′ terminus[33]. Both types of bias are difficult to quantify and correct. However, perhaps the most important limitation to existing *in vivo* RNA structure approaches is the challenge of analyzing low abundance RNA species. Not only do sequencing costs make their analysis prohibitive on a genome-wide scale, but input requirements for current low-throughput methods often necessitate *in vitro* transcription prior to structure profiling[15,16,30,34]. While *in vitro* RNA structure determination is valuable and informative, the barriers for *in vivo* investigation remain too high given the essential nature of these experiments and the vast number of low abundance RNA species, such as lncRNAs, whose structural conformations are of high scientific interest. Based on these limitations, we were motivated to develop a new RNA structure probing approach, called DMS-MaPseq, that would enable substantially more experimental investigation in the role of RNA structure *in vivo* than previous methods allowed.

FIGURES

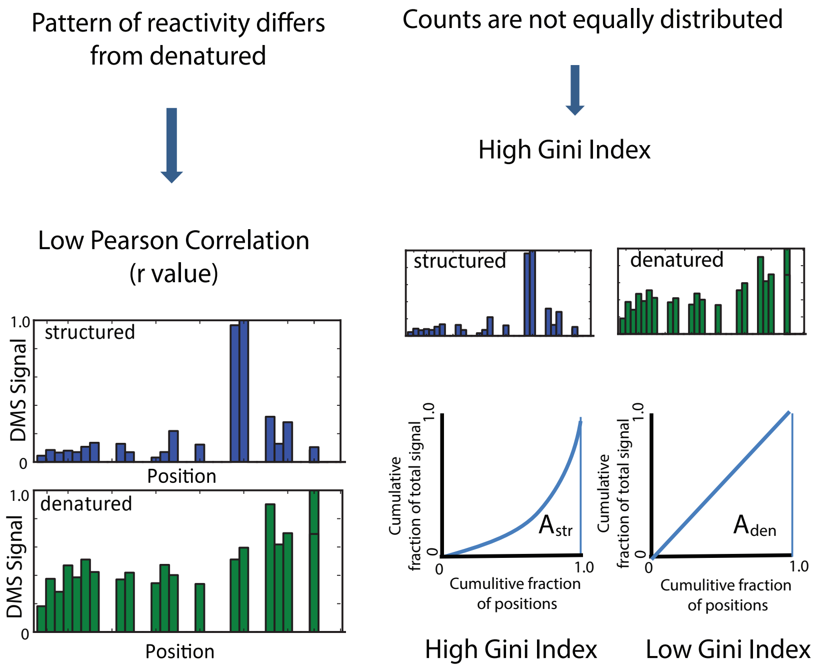

Pattern of reactivity differs from denatured
Low Pearson Correlation (r value)
Counts are not equally distributed
High Gini Index
1.0
0
DMS Signal
structured
Position
denatured
Cumulative fraction of total signal
Cumulitive fraction of positions
A_str
A_den
High Gini Index
Low Gini Index

Figure 1-1 | Schematic representaion of the two structure metrics used to define structured regions within mRNAs.

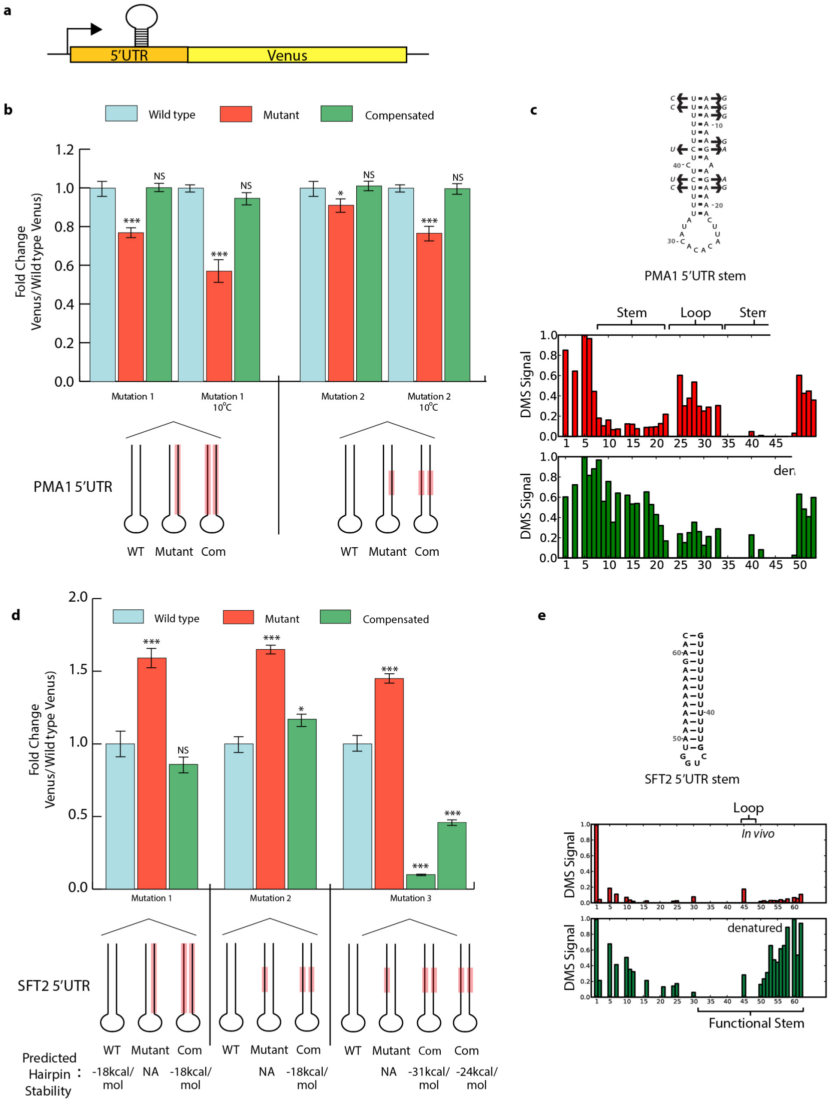
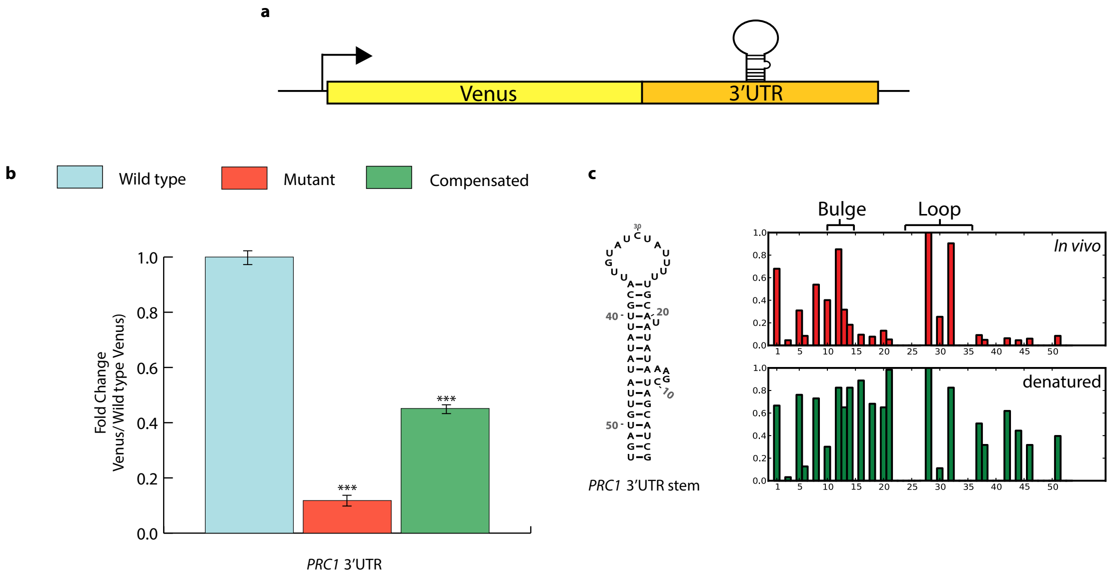


Figure 1-2 | Functional verification of novel 5'UTR structures *in vivo.* **a,** Putative 5'UTR stems were manipulated in the context of a Venus reporter *in vivo* **b,** *PMA1* 5'UTR structure was mutated and compensated twice with Venus reporter, differing in number and character of bases mutated. Mutation location shown in red on schematic. Reported p-values relative to wildtype Venus levels, calculated by two-sided t-test (p < .01, .001, and .0001 represent *, **, and *** respectively). For all graphs, Venus signal normalized to cell size before calculating fold change and data presented is from two biological and two technical replicates. Error bars represent SEM. **c,** Secondary structure of functional *PMA1* 5'UTR stem, with compensatory mutations (arrows) found in *S. paradoxus, S. mikatae, S. kudriavzevii,* and *S. bayanus*. Raw DMS signal shown below (position 1 = chrVII:482745). **d,** *SFT2* 5'UTR structure was mutated and compensated three times in Venus reporter system, differing in number, character, and location of bases mutated. Mutation location shown in red on schematic. Stem stability as predicated by mfold. Reported p-values relative to wild type Venus levels, also by two-sided t-test (p < .01, .001, and .0001 represent *, **, and *** respectively). Error bars represent standard deviation **c,** Secondary structure of functional *SFT2* 5'UTR stem. Position 1 = chrII:24023.



d

3’UTR	Chromosome and Position	Mutant Fold Change	Compensated Fold Change
SAM2	chr IV: 1,453,207-1,453,250	1.15 (p = 0.02)	1.46 (p = 1.3 x 10^{-5})
YDL085C-A	chr IV: 302,416-302,440	1.03 (p = 0.48)	0.99 (p = 0.79)
TSC3	chr II: 356,285-356,245	1.02 (p = 0.57)	0.94 (p = 0.05)

Figure 1-3 | Functional verification of novel *PRC1* 3’UTR structure *in vivo*. **a**, Putative 3’UTR stems were manipulated in the context of a Venus reporter in vivo, followed by Venus quantitation with flow cytometry. **b,** *PRC1* 3’UTR structure was mutated and compensated in Venus reporter system. For all data, reported p-values relative to wildtype Venus levels, calculated by two-sided t-test (p < .01, .001, and .0001 represent *, **, and *** respectively). Venus signal was normalized to cell size with fold change reported relative to Venus levels seen with the wild type stem. All results shown are derived from four measurements: two biological and two technical replicates. Error bars show standard deviation. **c,** Secondary structure of functional *PRC1* 3’UTR stem, shown with raw DMS signal for in vivo and denatured samples. Position 1 = chrXIII:863554. **d**,Weakly structured 3’UTRs *in vivo* were tested for function as in (b) but reveal little effectwhen mutated and no evidence for compensation.

SUPPLEMENTAL INFORMATION

	Stem Sequence (5’ to 3’)
PMA1 Wildtype	TTTTTTCTcTCTTTTatacacacattcAAAAGAaAGAAAAAA
PMA1 Mutant 1	TTTTTTCTcTCTTTTatacacacattcTTTTCTcTCTTTTTTT
PMA1 Compensated 1	AAAAAAGAaAGAAAAatacacacattcTTTTCTcTCTTTTTTT
PMA1 Mutant 2	TTTTTTCTcTCTTTTatacacacattcAAAttcatttAAAAA
PMA1 Compensated 2	TTTTTaagcgagTTTatacacacattcAAAttcatttAAAAA
SFT2 Wildtype	GTTTTTTTTTTTTGctggTAAAAAAAAAGAAC
SFT2 Mutant 1	CAAAAAAAAAAAATctggTAAAAAAAAAGAAC
SFT2 Compensated 1	CAAAAAAAAAAAATctggGTTTTTTTTTTTTG
SFT2 Mutant 2	GTTTTTTTTTTTTGctggTAAAttttttGAAC
SFT2 Compensated 2	GTTTaaaaaaTTTGctggTAAAttttttGAAC
SFT2 Mutant 3	GTTTTTTTTTTTTGctggTAAAccccccGAAC
SFT2 Compensated 3	GTTTggggggTTTGctggTAAAccccccGAAC
PRC1 Wildtype	GCTACGATcgaaATATAtACGTttttatctatgttACGTTATATATTGTAGT
PRC1 Mutant	TGATGTTAcgaaTATATtTGCAttttatctatgttACGTTATATATTGTAGT
PRC1 Compensated	TGATGTTAcgaaTATATtTGCAttttatctatgttTGCAATATATAGCATCG

Table 1-S1 | Sequences of functional structure mutations. 5’-3’ mRNA structure sequences are listed. Lowercase letters correspond to non-paired bases, found in bulges or loops within the stem. Mutated bases are highlighted in yellow.

MATERIALS AND METHODS

Media and Growth Conditions. Yeast strain BY4741 was grown in YPD at 30°C. For 10°C experiments, cells were grown to exponential phase by culturing for 72 hours at 10°C.

Functional UTR cloning. A fluorescent Venus reporter driven by a Nop8 promoter (chromosome XV:52262–53096) and *C. albicans ADH1* terminator was genomically integrated into yeast strain BY4741 at the *TRP1* locus (chromosome IV:461320–462280). Plasmids containing kanamycin resistance and the UTR of interest were made in a pUC18 plasmid backbone (Thermo Scientific). For the *PMA1* 5′ UTR, the entire 1-kb promoter region and 5′ UTR (chromosome VII: 482672–483671) was used. The pNop8 promoter was retained for the *SFT2* 5′ UTR investigation, with only the Nop8 5′ UTR replaced by the *SFT2* 5′ UTR. All 3′ UTRs were cloned to include > 100 base pairs after evidence of transcription ends. BY4741-Venus yeast were transformed using the standard technique of homologous recombination from a plasmid PCR product containing either a wild-type, mutant or compensated UTR. Successfully transformed yeast were identified by check PCR and subsequently sequenced to confirm the presence of only the desired mutations.

Mutagenesis in the endogenous PMA1 locus was done via the strategy described above for the PMA1 5′ UTR, except homologous recombination was targeted to the endogenous PMA1 locus and surrounding genomic region rather than to Venus. After sequencing to confirm the presence of only the desired mutations, PMA1 was carboxy-terminally tagged with Venus via PCR product from the pFA6a-link-yEVenus-SpHIS5 plasmid[35].

Flow cytometry. A saturated yeast culture was diluted 1:200 fold in minimal media and grown at 30°C for 6–8 hours before flow cytometry using a LSRII flow cytometer (Becton Dickinson) and 530/30 filter. Venus signal from each cell was normalized to cell size (Venus/side scatter) using Matlab 7.8.0 (Mathworks), and once normalized, all events (~20,000 per experiment) were averaged for a final Venus/side scatter value.

REFERENCES

1. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391,** 806–811 (1998).
2. Hannon, G. J. RNA interference. *Nature* **418,** 244–251 (2002).
3. Esteller, M. Non-coding RNAs in human disease. *Nat. Rev. Genet.* **12,** 861–874 (2011).
4. Rüegsegger, U., Leber, J. H. & Walter, P. Block of HAC1 mRNA translation by long-range base pairing is released by cytoplasmic splicing upon induction of the unfolded protein response. *Cell* **107,** 103–114 (2001).
5. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324,** 255–258 (2009).
6. Latrèche, L., Jean-Jean, O., Driscoll, D. M. & Chavatte, L. Novel structural determinants in human SECIS elements modulate the translational recoding of UGA as selenocysteine. *Nucleic Acids Res.* **37,** 5868–5880 (2009).
7. Chartrand, P., Meng, X. H., Singer, R. H. & Long, R. M. Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle in vivo. *Curr. Biol. CB* **9,** 333–336 (1999).
8. Jambor, H., Brunel, C. & Ephrussi, A. Dimerization of oskar 3′ UTRs promotes hitchhiking for RNA localization in the Drosophila oocyte. *RNA N. Y. N* **17,** 2049–2057 (2011).
9. Jambor, H., Mueller, S., Bullock, S. L. & Ephrussi, A. A stem-loop structure directs oskar mRNA to microtubule minus ends. *RNA N. Y. N* **20,** 429–439 (2014).
10. Winkler, W. C. & Breaker, R. R. Regulation of bacterial gene expression by riboswitches. *Annu. Rev. Microbiol.* **59,** 487–517 (2005).
11. Mercer, T. R. & Mattick, J. S. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* **20,** 300–307 (2013).

12. Lucks, J. B. *et al.* Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. U. S. A.* **108,** 11063–11068 (2011).
13. 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).
14. Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505,** 696–700 (2014).
15. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11,** 959–965 (2014).
16. Homan, P. J. *et al.* Single-molecule correlated chemical probing of RNA. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 13858–13863 (2014).
17. Spitale, R. C. *et al.* Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519,** 486–490 (2015).
18. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (2010).
19. Zheng, Q. *et al.* Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS Genet.* **6,** e1001141 (2010).
20. Underwood, J. G. *et al.* FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods* **7,** 995–1001 (2010).
21. Li, F. *et al.* Global analysis of RNA secondary structure in two metazoans. *Cell Rep.* **1,** 69–82 (2012).
22. Wan, Y. *et al.* Genome-wide measurement of RNA folding energies. *Mol. Cell* **48,** 169–181 (2012).

23. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505,** 706–709 (2014).

24. Kudla, G., Granneman, S., Hahn, D., Beggs, J. D. & Tollervey, D. Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 10010–10015 (2011).

25. Lu, Z. *et al.* RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell* **165,** 1267–1279 (2016).

26. Aw, J. G. A. *et al.* In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Mol. Cell* **62,** 603–617 (2016).

27. Sharma, E., Sterne-Weiler, T., O'Hanlon, D. & Blencowe, B. J. Global Mapping of Human RNA-RNA Interactions. *Mol. Cell* **62,** 618–626 (2016).

28. Wells, S. E., Hughes, J. M., Igel, A. H. & Ares, M., Jr. Use of dimethyl sulfate to probe RNA structure in vivo. *Methods Enzymol.* **318,** 479–493 (2000).

29. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127,** 4223–4231 (2005).

30. Mortimer, S. A. & Weeks, K. M. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.* **129,** 4144–4145 (2007).

31. Spitale, R. C. *et al.* RNA SHAPE analysis in living cells. *Nat. Chem. Biol.* **9,** 18–20 (2013).

32. Poulsen, L. D., Kielpinski, L. J., Salama, S. R., Krogh, A. & Vinther, J. SHAPE Selection (SHAPES) enrich for RNA structure signal in SHAPE sequencing-based probing data. *RNA* **21,** 1042–1052 (2015).

33. Aviran, S. & Pachter, L. Rational experiment design for sequencing-based RNA structure mapping. *RNA N. Y. N* **20,** 1864–1877 (2014).

34. Inoue, T. & Cech, T. R. Secondary structure of the circular form of the Tetrahymena rRNA intervening sequence: a technique for RNA structure analysis using chemical probes and reverse transcriptase. *Proc. Natl. Acad. Sci. U. S. A.* **82,** 648–652 (1985).

35. Sheff, M. A. & Thorn, K. S. Optimized cassettes for fluorescent protein tagging in Saccharomyces cerevisiae. *Yeast Chichester Engl.* **21,** 661–670 (2004).

CHAPTER TWO

Development of the DMS-MaPseq approach

INTRODUCTION

We sought to develop an *in vivo* and genome-wide approach that would overcome the limitations of truncation-based approaches by encoding DMS lesions as mutations instead, as has been recently described *in vitro* for individual RNA targets[1–3]. Such a mutational profiling (MaP) approach confers several useful features. These include the resolution of biases proximal to the information-encoding nucleotide and the analysis of multiple chemical modifications sites per molecule, opening up new possibilities for single-molecule RNA structure analysis such as the disambiguation of heterogeneous structure subpopulations *in vivo*. Finally, we reasoned that a MaP approach would make it possible to perform targeted amplification of low abundance RNA species while retaining a record of the sites of modifications. Yet mutation-based methods face a number of challenges to ensure the high signal and low background required for application on a genome-wide scale. A substantial increase in sequencing depth as a way to enhance signal over background is cost-prohibitive in genome-wide experiments. Additionally, given the particular utility of genome-wide experiments in the *de novo* discovery of functional RNA structures[4], any structural artifacts derived from background noise must be minimized.

Here we describe DMS-MaPseq, a novel RNA structure probing strategy that takes advantage of a high fidelity and processive thermostable group II reverse transcriptase (TGIRT) enzyme. We apply this technique to achieve the first mutation-based probing of RNA structure *in vivo*, both globally and for selected RNA species. DMS-MaPseq compares favorably to existing DMS-based approaches and delivers a low rate of insertions and deletions (indels), excellent detection of modifications, and low background error. We present the genome-wide application of DMS-MaPseq in both yeast and human cells, and we highlight a simple RT-PCR approach for targeted amplification in Chapter 3, focusing on RNA species inaccessible to previous techniques. For example, we apply DMS-MaPseq to examine mRNA structure during development in intact *D. melanogaster* ovaries establishing the suitability of DMS-MaPseq for probing RNA structure in animal tissue. Additionally, with DMS-MaPseq it is now possible to probe the structure of rare mRNA targets (<1 copy per cell) and to experimentally distinguish the RNA structures of different mRNA isoforms within the same physiological sample, demonstrated by our disambiguation of pre-mRNA structure from that of its mature, spliced counterpart in Chapter 4. Finally, we use DMS-MaPseq to reveal a functional 5′ structure in the human *FXR2* mRNA, which enables translation initiation at a non-canonical GUG codon[5]. With increased experimental versatility and improved data quality, DMS-MaPseq enables a far broader exploration of *in vivo* RNA structure and simultaneously offers an accessible technical solution to address structure-function hypotheses for virtually any RNA, regardless of abundance.

RESULTS

Unlike RT stop based approaches, the genome-wide strategy for *in vivo* DMS-MaPseq allows the detection of multiple modifications on a single RNA molecule enabling analyses of structurally heterogeneous populations. It also offers numerous advantages regarding data quality and experimental implementation (Fig. 2-1). For a genome-wide RNA structure experiment, we treat cells with high concentration of DMS, modifying up to 10% of open A/C bases. After total RNA extraction, random fragmentation, and the removal of ribosomal RNA, we ligate a 3′ adapter and reverse transcribe under conditions in which chemically modified bases are encoded as a mutation in the cDNA. A key advantage over RT stop methods is that the site of modification is not directly proximal to the fragment ends. Consequently, multiple modifications can be observed on a single cDNA fragment, and the data from DMS-MaPseq is inherently ratiometric (i.e., for any position the rate of modification is equal to the ratio of mutated reads to total reads) (Fig. 2-1b). This minimizes biases introduced during the library production, obviating the need for noisy computational corrections based on untreated or denatured control samples, which suffer from a combination of random and non-random background signal. While untreated or denatured DMS-MaPseq controls can still be useful in the discovery of endogenous mRNA modifications, uncharacterized single nucleotide polymorphisms, or as a negative control, it is no longer required for single nucleotide RNA structure calculations.

The accuracy of DMS-MaPseq depends critically on reverse transcription conditions that optimize the detection of DMS modifications while retaining high fidelity and processivity during cDNA synthesis. The thermostable group II intron reverse transcriptase (TGIRT) was recently adapted for molecular experimentation with these latter priorities in mind and notably produces mismatches at endogenous m^1A and m^3C tRNA residues—the exact methylation profiles of a DMS-modification[6,7]. Additionally, Superscript II with Mn^{2+} buffer (SSii/Mn^{2+}) has been used previously for the mutational read-through of DMS and SHAPE modification *in vitro* for the structural analysis of individual RNA molecules. To compare the suitability of these two enzymes for the *in vivo*, global DMS-MaPseq approach, we prepared genome-wide yeast libraries with both enzymes. One critical advantage of a chemical-based RNA structure probing approach is the single nucleotide resolution of the data collected. When encoding DMS modifications as mutations, mismatches inherently retain this advantage while insertions or deletions suffer from positional ambiguity when aligned across a homopolymeric stretch. As previously reported for SSii/Mn^{2+} conditions, RNA structure information is encoded either as mismatches or deletions, yet more than half of deletions must be discarded due to the above mapping ambiguity[3]. Indeed, we find that nearly a third of DMS-induced mutations from SSii/Mn^{2+} reverse transcription are insertions or deletions, compared to six percent for TGIRT (Fig. 2-2a). To assess the efficiency with which each reverse transcriptase detects DMS methylation, we used the two endogenous m^{1}A modifications on the yeast 25S rRNA as internal controls for DMS lesion detection. The ratiometric frequency of mismatches across three replicate experiments revealed that TGIRT detected these methylation sites with 84% and 46% frequency at the m^{1}A2142 and m^{1}A645 residues, respectively, placing a lower bound on the fraction of these residues that are endogenously modified. By contrast, SSii/Mn^{2+} yielded a mutation rate of only 54% and 3% at the same positions (Fig. 2-2b). Thus, the efficiency of modifications captured by SSii/Mn^{2+} may be as high as 65% at m^{1}A2142 but is less than 7% of the m^{1}A645 residues. This tendency of SSii/Mn^{2+} to underreport the DMS modification signal, and to do so in a context-dependent manner, could severely undermine its ability to robustly capture RNA structural information.

A valuable measure for the signal-to-noise ratio in sequencing data derived from DMS-modified samples is the enrichment of signal on adenines and cytosines[4] (Fig. 2-S1a). When the same source of DMS-modified RNA is subject to reverse transcription using either TGIRT or SSii/Mn^{2+}, we observed a far greater fraction of mismatches on A/Cs using TGIRT (93.5% versus 84%) (Fig. 2-2c). This high A/C signal in TGIRT data also exceeds that of our previously published DMS-seq strategy based on cDNA truncation events (90%)[4]. Additionally, the relative contributions of adenines and cytosines to the overall mismatch rate differ greatly between the various strategies. The low abundance of cytosines detected in DMS-seq indicates that truncation at cytosines is not robust[1]. Analysis of the mismatch nucleotide bias in DMS-seq reveals that 54% of mismatches occur on cytosines in a DMS-dependent manner, consistent with inefficient truncation at DMS-modified cytosines under those conditions (Fig. 2-S1 b,c). Notably, the signal on adenines is lower with SSii/Mn^{2+} than the other techniques, which suggests an underlying failure to robustly encode m^{1}A modifications consistent with the low signal detection on the endogenous rRNA residues (Fig. 2-2b). In summary, DMS-MaPseq with the TGIRT enzyme provides higher signal to noise and a more robust capture of both modified cytosines and adenines compared to SSii/Mn^{2+} or to the previously published truncation-based strategy.

We also assessed the *in vivo* DMS signal derived from these mutational profiling methods for a known positive control structure in the yeast *RPS28B* 3′ UTR[8] (Fig. 2-2d). While both RT protocols produced excellent signal at unpaired A/C residues in this structure, the SSii/Mn^{2+} data revealed high background signal on certain G/U residues, suggesting a propensity for non-random errors in cDNA synthesis that would adversely affect data quality. This higher background error for SSii/Mn^{2+} is also reflected in the genome-wide frequency of mutations and indels on matched untreated and DMS-treated RNA (Supplementary Fig. 1d). These observations of high background with Mn^{2+} buffer are consistent with its historical use in deliberate mutagenesis during oligonucleotide synthesis[9]. As previously mentioned, genome-wide techniques necessitate a stringent signal-to-noise ratio, so we used replicates to assess the reproducibility of the RNA structure signal across yeast transcriptome regions as measured by *r* value and the Gini index difference, two established metrics for RNA structure determination which measure the similarity in pattern and evenness of data distribution, respectively[4] (Fig. 2-2e). This analysis reveals a stronger reproducibility between data generated by TGIRT reverse transcription over SSii/Mn^{2+} and is consistent with our observations of high background noise in the latter approach. Due to high DMS signal detection and low background error observed across many quality control metrics, we chose the TGIRT enzyme for all further DMS-MaPseq experimentation and method development.

Global analysis of DMS-MaPseq data

When DMS lesions are detected by truncation, only the most 3′ DMS modification on an RNA fragment will be detected and information from additional DMS modifications is lost. For this reason, DMS treatment conditions must be carefully titrated to avoid improper hit kinetics and 5′ signal decay[10]. This effect is illustrated by the lack of DMS-seq signal immediately 5′ of an endogenous m^1A residue in denatured yeast 25S rRNA (Fig. 2-3a). However, denatured DMS-MaPseq data at the same rRNA location shows no drop in DMS signal, confirming the TGIRT enzyme can encode multiple DMS lesions in a short sequence space with no loss of signal from neighboring modifications. Additionally, negative control bases in the yeast rRNA (i.e. those known to be involved in stable secondary structure) fall overwhelmingly into the lowest bin of structure reactivity in DMS-MaPseq data, confirming the low background noise observed previously (Fig. 2-3b) and exhibiting an improvement over published DMS-seq data[4].

We also collected a genome-wide DMS-MaPseq dataset from the *in vivo* DMS treatment of human embryonic kidney (HEK) 293T cells, with a sequencing depth of ~200 million uniquely mapped reads, and we confirm the excellent agreement of our data with the *XBP1* positive control structure[11] (Fig. 2-S2a). Often, GC content is invoked as an indicator for RNA structure, so we investigated this relationship across human transcriptome regions, plotting GC content against the Gini index derived from our DMS-MaPseq data (Fig. 2-3c). A small correlation (r = 0.32) exists between these two metrics, but there is a clear difference in trends when separating RNA classes. Overall, coding regions have lower GC content and their RNA appears less structured, as we demonstrated previously, yet the lack of structure is more pronounced than would be predicted by GC content alone. Non-coding RNA regions, however, which include UTRs and all classes of mammalian ncRNAs, are more structured than CDS regions of comparable GC content. Interestingly, the biggest outliers are snoRNAs and snRNAs, which contain a relatively low fraction of GCs but are amongst the most structured regions analyzed, possibly due to their stabilization by protein binding.

FIGURES

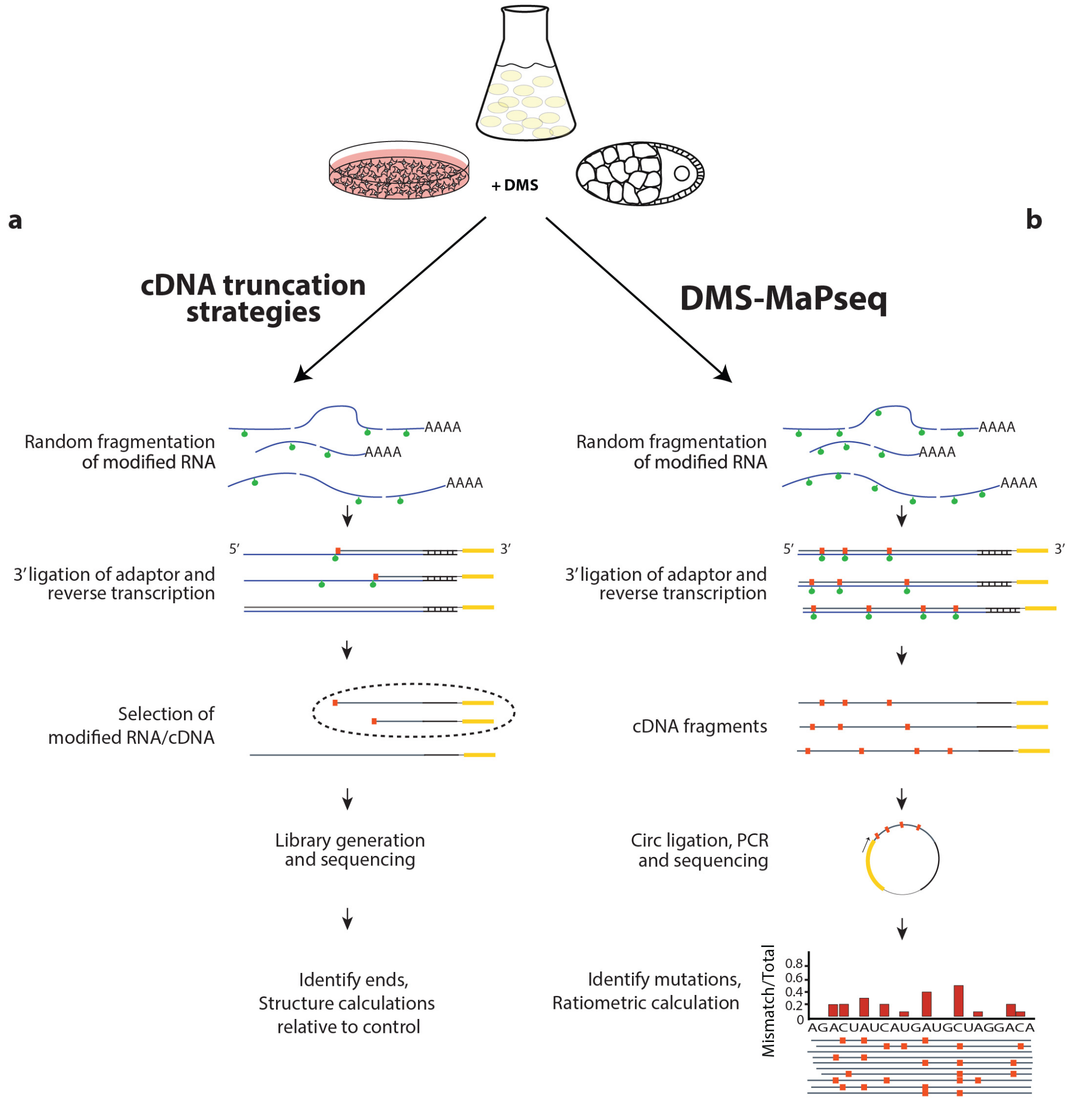
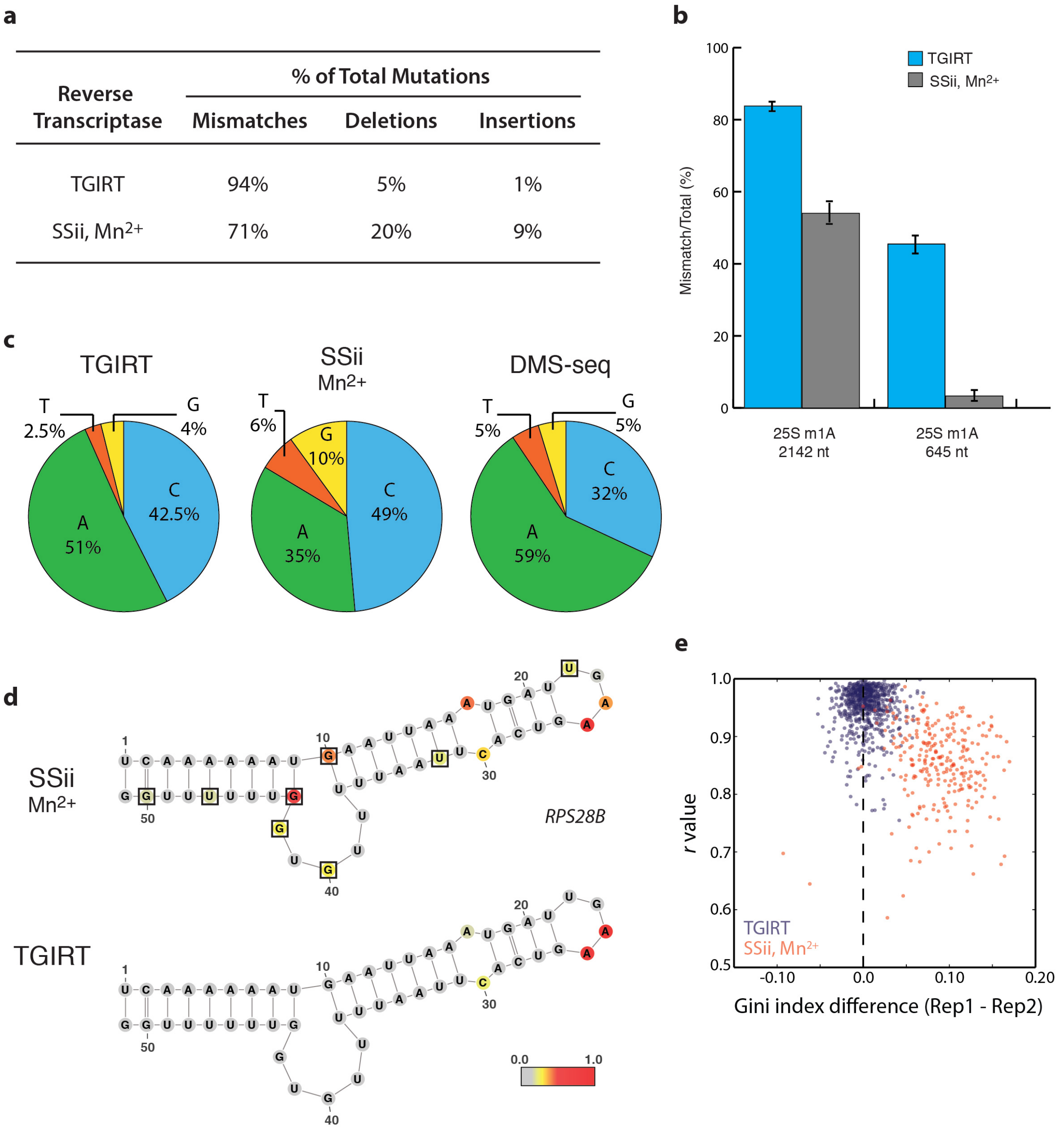

+ DMS
a
b
cDNA truncation strategies
DMS-MaPseq
Random fragmentation of modified RNA
AAAA
AAAA
AAAA
Random fragmentation of modified RNA
AAAA
AAAA
AAAA
5′
3′
3′ ligation of adaptor and reverse transcription
5′
3′
3′ ligation of adaptor and reverse transcription
Selection of modified RNA/cDNA
cDNA fragments
Library generation and sequencing
Circ ligation, PCR and sequencing
Identify ends, Structure calculations relative to control
Identify mutations, Ratiometric calculation
Mismatch/Total
0.8
0.6
0.4
0.2
0
AGACUAUCAUGAUGCUAGGACA

Figure 2-1 | Sequencing library generation for RNA structure probing techniques. Schematic of library preparation strategies for cDNA truncation approaches **(a)** and for DMS-MaPseq **(b)**, which has a higher DMS modification level, no selection for modified molecules, and ligations that are no longer directly proximal to the structure information-containing positions. The structure signal for DMS-MaPseq is inherently ratiometric and calculated per nucleotide as the number of mismatches divided by base sequencing depth. 5′ to 3′ orientation noted relative to RNA fragment.



a

Reverse Transcriptase	% of Total Mutations		
	Mismatches	Deletions	Insertions
TGIRT	94%	5%	1%
SSii, Mn^{2+}	71%	20%	9%

Figure 2-2 | TGIRT enzyme delivers higher signal and lower background for DMS-MaPseq. **a,** Distribution of mutation type generated by SSII/Mn^{2+} or TGIRT reverse transcription from *in vivo* DMS-treated yeast mRNA. **b,** Endogenous m^1A modifications in yeast 25S rRNA transcript reveal superior modification detection with TGIRT. Average percent modification detected at the position across three DMS-treated replicates with error bars representing standard deviation. **c,** Nucleotide composition of mismatches from TGIRT or SSII/Mn^{2+} approaches. **d,** Yeast *RPS28B* mRNA positive control structure with nucleotides colored by DMS reactivity *in vivo*. Black boxes outline G/U bases with high background signal. DMS reactivity was calculated as the average ratiometric DMS signal per position across two replicates normalized to the highest number of reads in displayed region, which is set to 1.0. **e,** Genome-wide DMS-MaPseq replicates compared by Pearson's *r* value and Gini index for yeast mRNA regions.

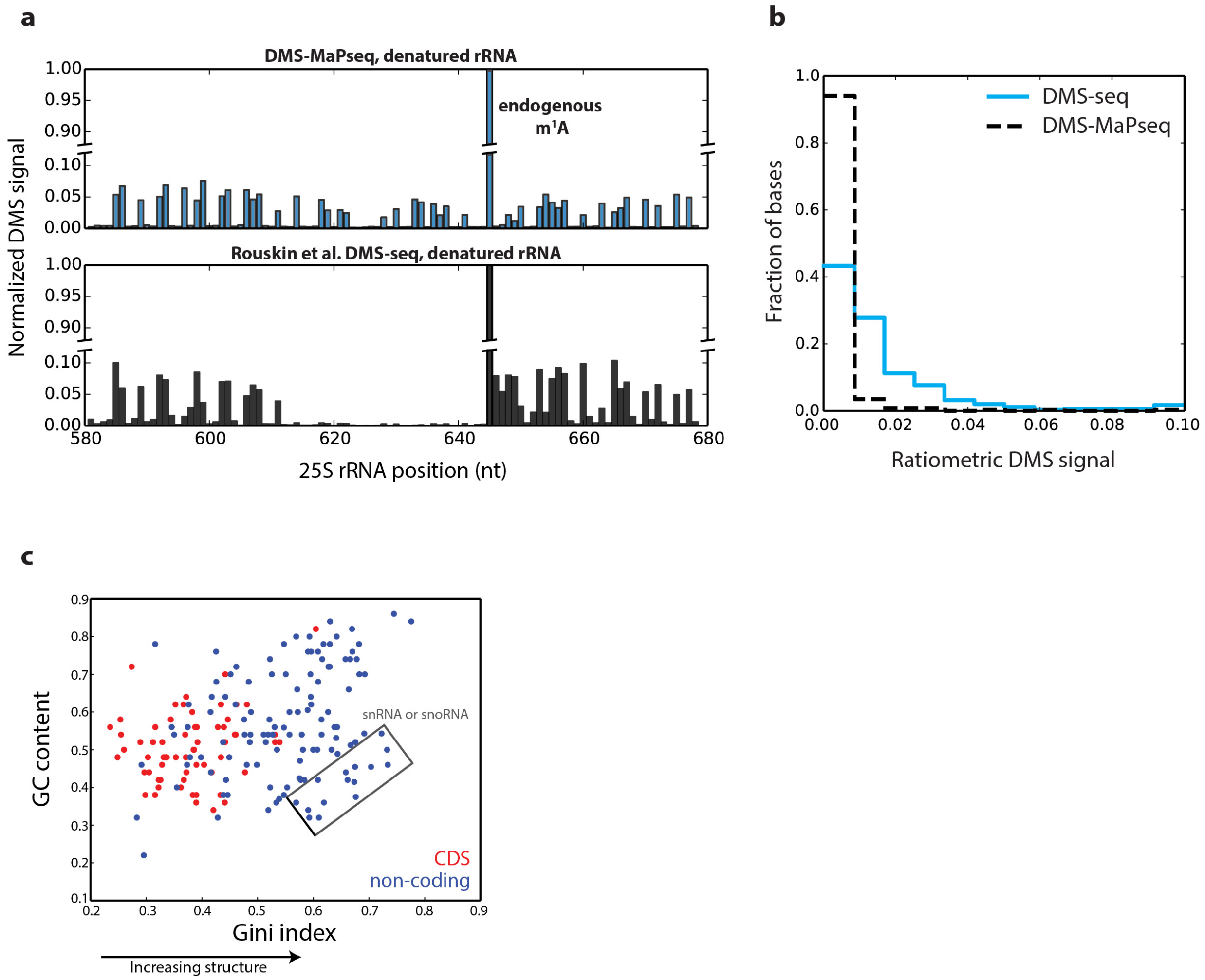
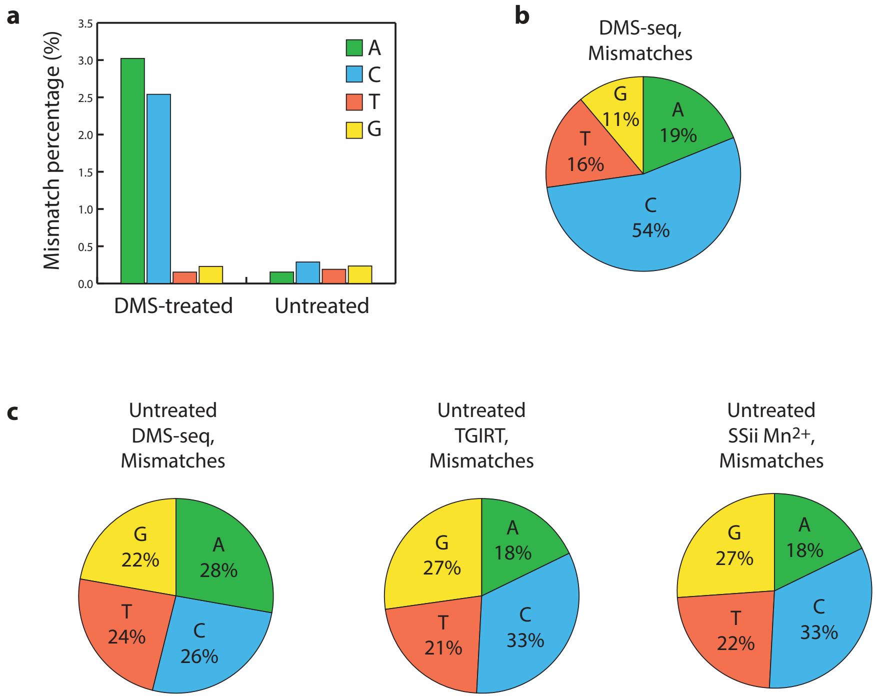


Figure 2-3 | Global analysis of *in vivo* DMS-MaPseq data. **a,** Signal decay observed after endogenous m^{1}A modification at position 642 in the yeast 25S rRNA in DMS-seq, but not in DMS-MaPseq. **b,** Histogram of ratiometric reactivity for negative control bases in the yeast 18S rRNA. **c,** Coding regions in genome-wide HEK 293T data appear less structured than ncRNA regions, even at comparable GC content.

SUPPLEMENTAL INFORMATION



d

	Mismatch Frequency (%)		Indel Frequency (%)	
Library Type	**Untreated**	**DMS-treated**	**Untreated**	**DMS-treated**
TGIRT	0.21%	1.50%	0.02%	0.10%
SSii, Mn^{2+}	0.26%	1.09%	0.08%	0.44%
DMS-seq	0.33%	0.51%	0.03%	0.03%

Figure 2-S1 | Mutations produced by reverse transcription on *in vivo* DMS-treated and untreated templates. **a,** Total mismatch percentage on each nucleotide from *in vivo* DMS-MaPseq with TGIRT on yeast mRNA. **b,** Nucleotide composition of mismatches in DMS-seq from Rouskin et al. for *in vivo* DMS-treated yeast mRNA, revealing a preference to generate mismatches on cytosines. **c,** Nucleotide composition of mismatches as detected by existing RNA structure probing approaches for untreated yeast mRNA, revealing no strong mismatch biases independent of DMS modification. **d,** Mutation frequency from DMS-treated and untreated yeast mRNA templates, derived from the same RNA source for TGIRT and SSii/Mn^{2+} data. Mutation frequency was calculated as the number of mismatches or indels detected via sequencing divided by the total number of bases sequenced.

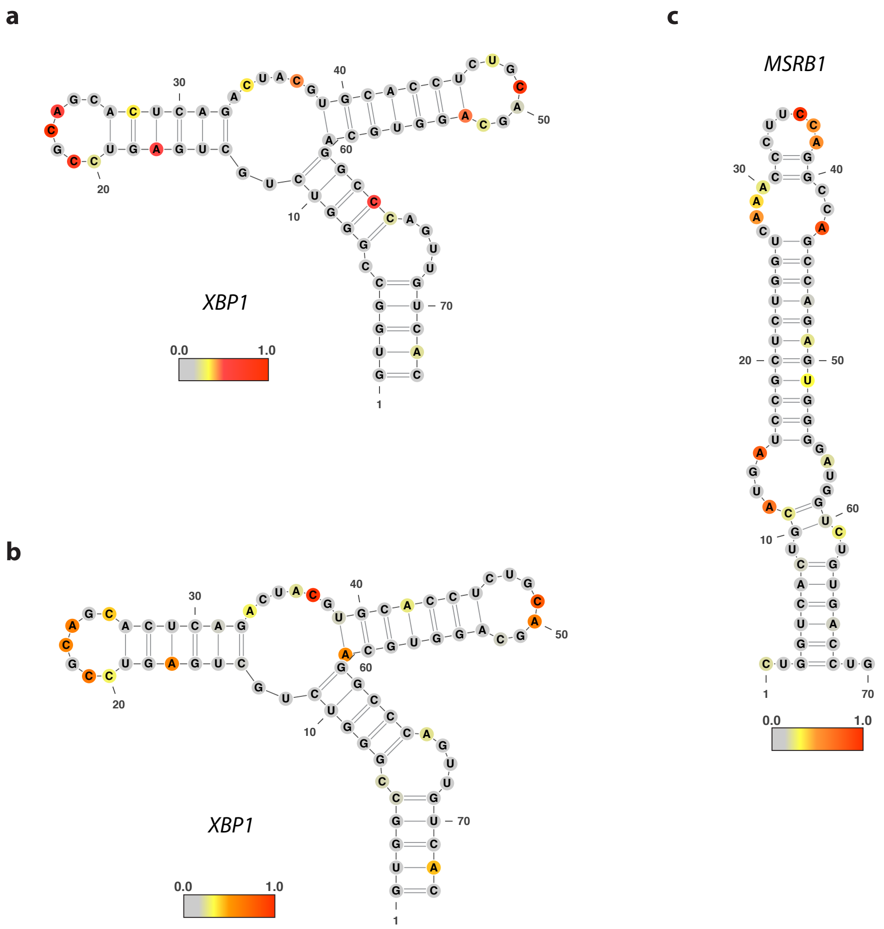


Figure 2-S2 | Positive control RNA structures from *in vivo* DMS treatment of HEK 293Ts. **a,** *XBP1* mRNA positive control structure with nucleotides colored by DMS reactivity from genome-wide DMS-MaPseq. **b, c,** *XBP1* and *MSRB1* mRNA positive control structures with nucleotides colored by DMS reactivity from target-specific DMS-MaPseq. DMS reactivity calculated as the ratiometric DMS signal per position normalized to the highest number of reads in displayed region, which is set to 1.0.

Table 2-S1.

Primers used in this chapter.

name	purpose	sequence (5' to 3')
Linker 2	3' Cloning adaptor for RNA footprints	5rApp/CACTCGGGCACCAAGGA/3ddC
oCJ200-link2	Primer for reverse transcription of sequencing libraries	5'/5phos/GATCGTCGGACTGTAGAACTCTGAACCTGTCG/iSp18/CAAGCAGAAGACGGCATACGAGATTCCTTGGTGCCCGAGTG
oNTI231	Amplification of sequencing libraries, paired with indexing primer	caagcagaagacggcatacga
Indexing primer with 6bp TruSeq index		aatgatacggcgaccaccgagatctacacgatcggaagagcacacgtctgaactccagtcacNNNNNNcgacaggttcagagttc
oNTI202	Read1 sequencing primer	CGACAGGTTCAGAGTTCTACAGTCCGACGATC

MATERIALS AND METHODS

Media and growth conditions. Yeast strain BY4741 was grown in YPD at 30°C. Saturated cultures were diluted to OD_{600} of ~0.09 and grown to a final OD_{600} of 0.5-0.7 at the time of DMS treatment. HEK 293T cells were grown in DMEM medium with high glucose, supplemented with glutamine, pyruvate, non-essential amino acids, and 10% FBS, and cells were treated with DMS at ~80% confluence.

Dimethyl sulfate (DMS) modification. For *in vivo* DMS modification in yeast, 15 ml of exponentially growing yeast were incubated with 750 µl DMS (Sigma) for 4 min at 30°C. DMS was quenched by adding a 30 ml stop solution comprised of 30% beta-mercaptoethanol (from a 14.2 M stock) and 50% isoamyl alcohol, after which cells were quickly put on ice, collected by centrifugation at 3,500 x g at 4°C for 4 min, and washed with 10 ml 30% BME solution. Cells were then resuspended in 0.6 ml total RNA lysis buffer (6 mM EDTA, 45 mM NaOAc pH 5.5), and total RNA was purified with hot acid phenol (Ambion) and EtOH precipitation. Ribosomal RNA was depleted using RiboZero (Epicentre), either directly after RNA extraction or post-ligation in the genome-wide library preparation. Denatured RNA structure samples were treated as in DMS-seq[4]. For HEK 293T cells, 15 cm[12] plates with 15 ml of media were treated with the addition of 300 µl DMS and incubation at 37°C for 4 min. Media/DMS was decanted, and plates were washed twice in 30% BME (v/v). Cells were resuspended in Trizol, and RNA isolated according to manufacturer protocol.

Library generation, genome-wide DMS-MaPseq. Sequencing libraries were prepared with a modified version of the protocol used for DMS-seq[4]. Specifically, 10 µg of DMS-treated total RNA was denatured for 2 min at 95°C, then fragmented at 95°C for 2 min in 1X RNA Fragmentation Reagent (Zn^{2+} based, Ambion). The reaction was stopped with 1x Stop Solution

(Ambion) and quickly placed on ice. The fragmented RNA was run on a 6% TBU (Tris Borate Urea) polyacrylamide gel for 45 min at 150 V. A blue light (Invitrogen) was used for gel imaging, and RNA fragments of 100-170 nucleotides in size were excised, depleting small ncRNA contaminants of <100 nucleotides (tRNAs, snoRNAs). Gel extraction was performed by crushing the purified gel piece and incubating in 300 µl 300 mM NaCl at 70°C for 10 min with vigorous shaking. The RNA was then precipitated by adding 2 µl GlycoBlue (Invitrogen) and 3x volume (900 µl) 100% EtOH, incubating on dry ice for 20 min and spinning at 20k x g for 45 min at 4°C. The samples were then resuspended in 7 µl 1X CutSmart buffer (NEB) and the 3′ phosphate groups left after random fragmentation were resolved by adding 1.5 µl rSAP (NEB), 1 µl of SUPERase Inhibitor (Ambion) and incubating at 37°C for 1 hour. After heat inactivation of the phosphatase at 65°C for 5 min, the samples were then directly ligated to 25 pmol of miRNA cloning linker-2 (IDT) by adding 2 µl T4 RNA ligase2, truncated K227Q (NEB), 1 µl 0.1M DTT, 6.5 µl 50% PEG, 1 µl 10X T4 RNL2 buffer, and incubating for 2 hours at 25°C. Reactions were purified by EtOH precipitation (as above), and excess linker was degraded for 1 hour at 30°C in a 20 µL reaction of 1x RecJ buffer, 1 µl SUPERase Inhibitor, 1 µl 5′ Deadenylase (Epicentre), and 1 µl RecJ exonuclease (Epicentre). Ribosomal RNA was depleted using RiboZero (Epicentre), with a final incubation of 5 min at 40°C, instead of 50°C as recommended in the commercial protocol, and purified by EtOH precipitation. Reverse transcription was performed in a 10 µl volume with 1 pmol oCJ200-link2. To begin, a mixture of RNA/primer/buffer was incubated at 80°C for 2 min to denature the template, then returned to ice for the addition of SUPERase Inhibitor (Ambion), DTT, dNTPs, and RT enzyme to generate the final reaction conditions. For reverse transcription using Superscript II with Mn^{2+} buffer, we followed the exact published reactions conditions for mutational profiling[1] [0.5 mM dNTPs, 50 mM Tris-HCl pH 8.0, 75 mM KCl, 6 mM $MnCl_2$, and 10 mM DTT] and allowed the reaction to proceed for 2-3 h at 42°C with 100U of SuperScript II (Invitrogen). Due to potential pausing of the TGIRT at modification sites, this long incubation time facilitates readthrough of multiple modifications per

RNA fragment. For the TGIRT reverse transcription, a 5 min incubation at room temperature followed the initial denaturation, and the RT reaction proceeded for 1.5 h at 57°C with 100 U TGIRT-III enzyme (InGex) and the following reaction conditions: 1 mM dNTPs, 5 mM freshly prepared DTT (Sigma-Aldrich), 10 U SUPERase Inhibitor, 50 mM Tris-HCl pH 8.3, 75 mM KCl, 3 mM $MgCl_2$. After reverse transcription, 1 µl of 5 N NaOH was added and the reaction incubated for 3 min at 95°C to degrade the RNA, followed by EtOH precipitation and gel purification to remove excess RT primer. Finally, cDNAs were circularized using CircLigase (Epicentre), and Illumina sequencing adapters and indexes were introduced by 9-13 cycles of PCR using Phusion HF Polymerase (NEB), oNTI231, and indexing primers with TruSeq 6 bp indices. Libraries were sequenced with oNTI202 in 50 nt single-end reads on the HiSeq4000 (Illumina). See primer sequences in Table 2-S1.

Sequencing alignment and analysis. Raw fastq files were stripped of linker sequences and filtered for quality using the FASTX-Toolkit Clipper and Quality Filter functions, respectively, requiring that 80% of sequenced bases have a quality score >25 (http://hannonlab.cshl.edu/fastx_toolkit/). Reads were aligned using Tophat v2.1.0 with bowtie2 with the following settings for a 50 nt sequencing run: --no-novel-juncs -N 5 --read-gap-length 7 --read-edit-dist 7 --max-insertion-length 5 --max-deletion-length 5 -g 3. All non-uniquely aligned reads were then removed. Sequencing data was aligned against the *Saccharomyces cerevisiae* assembly R64 (UCSC: sacCer3) downloaded from the Saccharomyces Genome Database on February 8, 2011 (SGD, www.yeastgenome.org) or against the longest human RefSeq isoforms (hg19). Due to empirically determined mutation enrichment from non-template addition, we trimmed 2 and 5 nucleotides from the 5′ end of each read for TGIRT and SSii/Mn^{2+} generated libraries, respectively. Mismatches located within 3 nucleotides of an indel were also discarded for future analysis. The ratiometric DMS signal was calculated for each nucleotide as # mismatches / sequencing depth. Genome-wide yeast DMS-MaPseq data was collected and sequenced with two biological replicates for each SSii/Mn^{2+} and TGIRT, untreated and *in vivo* DMS-treated libraries. For each library variation, we collected a combined total of 90 to 200 million uniquely mapped reads between yeast replicates and 200 million for HEK 293T cells.

Secondary structure models. Novel secondary structure models were generated using constraints derived DMS-MaPseq data using RNAfold[13]. DMS-MaPseq reactivities were overlaid on structure models using VARNA (http://varna.lri.fr/)[14].

HEK 293T Gini index calculations. UTR and coding regions were defined by RefSeq coordinates, and we analyzed 50 nt windows beginning at the annotated transcription start site. After requiring a minimum number of 100 total reads at A/Cs and >20x mismatch coverage for each window, we also discarded any windows with evidence for endogenous modifications (>15% mutation rate). The Gini index was calculated only for A/C bases, as done previously[4].

REFERENCES

1. Homan, P. J. *et al.* Single-molecule correlated chemical probing of RNA. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 13858–13863 (2014).
2. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11,** 959–965 (2014).
3. Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A. & Weeks, K. M. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat. Protoc.* **10,** 1643–1669 (2015).
4. 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).
5. Fields, A. P. *et al.* A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol. Cell* **60,** 816–827 (2015).
6. Mohr, S. *et al.* Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA N. Y. N* **19,** 958–970 (2013).
7. Katibah, G. E. *et al.* Broad and adaptable RNA structure recognition by the human interferon-induced tetratricopeptide repeat protein IFIT5. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 12025–12030 (2014).
8. Badis, G., Saveanu, C., Fromont-Racine, M. & Jacquier, A. Targeted mRNA degradation by deadenylation-independent decapping. *Mol. Cell* **15,** 5–15 (2004).
9. Beckman, R. A., Mildvan, A. S. & Loeb, L. A. On the fidelity of DNA replication: manganese mutagenesis in vitro. *Biochemistry (Mosc.)* **24,** 5810–5817 (1985).
10. Aviran, S. & Pachter, L. Rational experiment design for sequencing-based RNA structure mapping. *RNA N. Y. N* **20,** 1864–1877 (2014).

11. Hooks, K. B. & Griffiths-Jones, S. Conserved RNA structures in the non-canonical Hac1/Xbp1 intron. *RNA Biol.* **8,** 552–556 (2011).

12. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).

13. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol. AMB* **6,** 26 (2011).

14. Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinforma. Oxf. Engl.* **25,** 1974–1975 (2009).

CHAPTER THREE

DMS-MaPseq enables target-specific amplification for *in vivo* RNA structure investigation of any RNA of interest

INTRODUCTION

Certain RNA structure experiments cannot be addressed adequately by a genome-wide approach. For example, low abundance mRNAs do not receive sufficient sequencing coverage in genome-wide experiments to make robust conclusions about their structure. Plotting the cumulative *r* value distribution for mRNA regions between *in vivo* DMS-MaPseq replicates in yeast reveals that an average mismatch coverage depth of greater than 20x greatly improves data reproducibility (Fig. 3-1a). However, for genome-wide HEK 293T DMS-MaPseq data with 50, 100, or 200 million uniquely mapped reads, only a limited fraction of genes—0.006, 0.009, and 0.03, respectively—pass this 20x coverage threshold (Fig. 3-1b). Even when extrapolated to an exorbitant sequencing depth of 1 billion uniquely mapped reads, many human genes (78%) have insufficient coverage. Because many RNAs of high biological importance and interest are lowly expressed (such as lncRNAs or those implicated in Mendelian disorders), a target-specific approach to assess the *in vivo* RNA structure of modest or low expression RNA targets would be hugely valuable.

RESULTS

DMS-MaPseq for specific or low abundance RNA targets

To probe the *in vivo* structure of low abundance mRNAs, we developed and validated a simple targeted RT-PCR implementation of DMS-MaPseq (Fig. 3-1c). In DMS-MaPseq, the position of the chemical modification is imprinted in the cDNA as a mutation, and this information is retained through rounds of PCR amplification. Similar to the genome-wide approach, targeted DMS-MaPseq begins with the *in vivo* modification of RNA, followed by total RNA extraction. After DNase treatment and an rRNA depletion step, we reverse transcribe using the TGIRT enzyme and gene-specific primers, which can be used in combination to amplify multiple RNA species in a single reaction. Directly after cDNA synthesis, gene-specific PCR primers amplify the RNA region of interest, followed by Nextera tagmentation to fragment the dsDNA and add adaptors for sequencing.

To assess the quality of data derived from this targeted DMS-MaPseq approach, we examined the structure signal for known RNA structures. In our most stringent test for data quality, we plotted an ROC curve to assess the agreement of 18S rRNA DMS-MaPseq data with the published yeast crystal structure model[1] and observed an excellent agreement when data was collected by either our genome-wide or targeted approach (Fig. 3-1d). In addition to the ROC curve, we confirmed that the *HAC1* 3′ UTR secondary structure is supported by DMS signal derived from this targeted DMS-MaPseq approach, as is the *RPS28B* 3′ UTR positive control structure[2,3] (Fig. 3-1 e,f). We also applied targeted DMS-MaPseq to the human *XBP1* and *MSRB1* RNAs and, similarly, find that DMS-MaPseq data is in excellent concordance with their known structure models[4,5] (Fig. 3-S1 b,c).

To remove PCR amplification biases for quantitative applications involving low input material, we also developed a variation of targeted DMS-MaPseq that tags each RNA molecule with a unique molecular index (UMI). Using a gene-specific RT primer with a 5′ overhang comprised of an N_{10} random index and defined PCR primer binding site, each cDNA is labeled with a UMI (Fig. 3-S1a). After gene-specific PCR amplification and limited-cycle second PCR to add sequencing adaptors and indexes, the PCR amplicon is sequenced on a MiSeq for a read length specified by the size of the region of interest. Unique reads can then be easily isolated based on their specific UMI and DMS mutation profile. In addition to future quantitative applications for this UMI-based data, it assists the structure profiling of low-abundance mRNAs by guarding against “jackpotting” effects, i.e. when many copies of a single molecule take over a population during PCR amplification. The *ASH1* and *SFT2* yeast mRNAs are lowly expressed and host functional RNA structures in their 3′ and 5′ UTRs, respectively, serving as positive controls for DMS signal detection utilizing a UMI. Indeed, both controls show DMS modification profiles in excellent agreement with the known secondary structure models[6,7] (Fig. 3-S1 b,c). In summary, our target-specific DMS-MaPseq approach generates excellent RNA structure data, is easy and cost-effective to implement, and can be adapted for highly quantitative applications through the addition of unique molecular index.

FIGURES

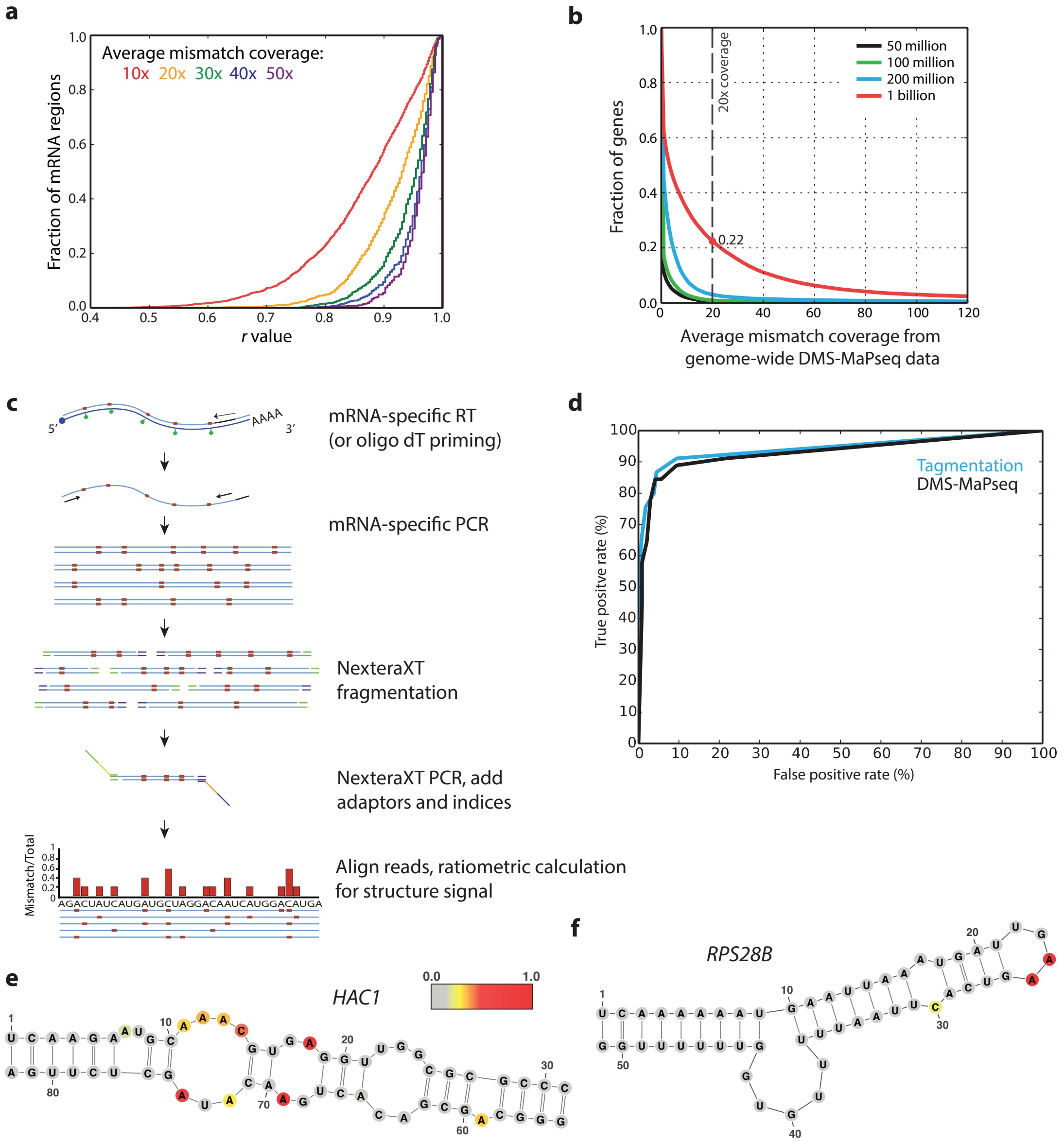


Figure 3-1 | DMS-MaPseq enables *in vivo* RNA structure probing for specific RNA targets. **a,** Cumulative histogram of Pearson's *r* values between yeast mRNA regions in DMS-MaPseq replicates at varied depths of average mismatch coverage. **b,** Fraction of genes exceeding the minimum average mismatch coverage of 20x in genome-wide human HEK 293T DMS-MaPseq data with varied sequencing depths. **c,** Schematic for targeted RNA structure probing via target-specific RT-PCR and NexteraXT tagmentation. **d,** ROC curve for DMS signal on yeast 18S rRNA using ratiometric data from target-specific tagmentation approach and from genome-wide DMS-MaPseq. **e, f,** Yeast *HAC1* (e) and *RPS28B* (f) 3′ UTR mRNA positive control structures from target-specific priming with nucleotides colored by DMS reactivity *in vivo*. DMS reactivity calculated as the ratiometric DMS signal per position normalized to the highest number of reads in displayed region, which is set to 1.0.

SUPPLEMENTARY INFORMATION

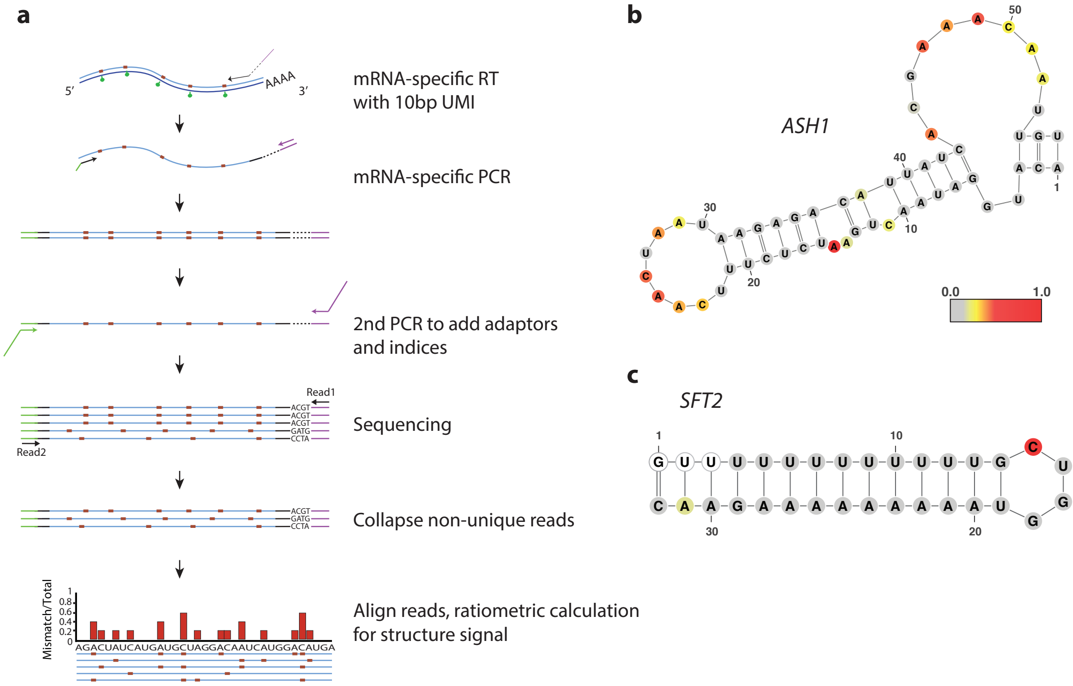


Figure 3-S1 | Targeted amplification of low-abundance RNA targets using a unique molecular index. **a,** Schematic for targeted RNA structure probing via gene-specific RT-PCR using a unique molecular index (UMI) on the RT primer. **b, c,** Yeast *ASH1* (b) and *SFT2* (c) mRNA positive control structures from target-specific UMI approach with nucleotides colored by DMS reactivity *in vivo*. DMS reactivity calculated as the ratiometric DMS signal per position normalized to the highest number of reads in displayed region, which is set to 1.0. Uncolored nucleotides had no data collected.

Table 3-S1.

Primers used in this chapter.

name	purpose	sequence (5' to 3')
oMZ282	Reverse transcription and reverse primer (1st PCR) for targeted amplification with a unique molecular index	GCAGCGACAGGTTCAGAGTTCTACAGTCCGACGATC – $(N)_{10}$ – gene-specific primer
oMZ283	Forward primer (1st PCR) for targeted amplification with a unique molecular index	CTGAACCGCTCTTCCGATCT–gene-specific primer
oMZ284	Forward primer (2nd PCR) for targeted amplification with a unique molecular index	CAAGCAGAAGACGGCATACGAGACGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
oMZ285	*ASH1* RT primer	oMZ282 + TTTGTAGTTTATTTAGCACAGACAAGGAGAGAAATGT
oMZ286	*ASH1* Fwd PCR primer	oMZ283 + TGGAATAGACAAAGAATTCGTCCCCGAACGCAACT
oMZ287	*SFT2* RT primer	oMZ282 + CGAGTTCTGCTGTCTTGTTTGATTCCATCG
oMZ288	*SFT2* Fwd PCR primer	oMZ283 + GAGTAGGACTTAGATTACCTGTATTGTCTGCAGTTGCGTT
oMZ289	*RPL14A* RT/Rev PCR primer, intron	ACGATGGAAGGCATGGTTTAATATTTGAGGAAACATGG
oMZ290	*RPL14A* RT/Rev PCR primer, exon	AGCCTTAGCCAAAGCCTTCTTGACAGTGTA
oMZ291	*RPL14A* Fwd PCR primer	TGTCCACCGATTCTATTGTCAAGGCTTCTAACTGG
oMZ292	*RPL31B* RT/Rev PCR primer, intron	AAGGTGGAACTAAAGCATCACGCCAAAAACATCG
oMZ293	*RPL31B* RT/Rev PCR primer, exon	CGTACCCCGAAAGCAGCTCTGTTTGTGTAAT
oMZ294	*RPL31B* Fwd PCR primer	TGCACGAGCAGATAATCCAAAGTACTTGAAAATGGCC

MATERIALS AND METHODS

Library generation, targeted DMS-MaPseq. After culturing, *in vivo* DMS treatment, and total RNA extraction as outlined in Chapter 2, 5 µg of total RNA was DNase-treated for 30min at 37°C in 1x TURBO DNase buffer with 1 µl TURBO DNase enzyme (Thermo Fisher Scientific). Reactions were desalted using RNA Clean & Concentrator-5 columns (Zymo Research), and rRNA was depleted using RiboZero (Epicentre) or with RNase H for *D. melanogaster* and HEK 293T samples, implemented with slight modifications to the published protocol[8]. For the RNase H protocol, briefly, 5 µg of total RNA was depleted of small RNA species with a Zymo RNA Clean & Concentrator-5 column, retaining RNA >200 nt per manufacturer instructions. RNase H subtraction was performed by adding 5 µg of published subtraction oligos[8] in a total volume of 30 µl in 1X Hybridization Buffer (200 mM NaCl, 100 mM Tris pH 7.5). The mixture was incubated at 68°C for 1 min, and the temperature was ramped down at a rate of 1°C / min down to 45°C. $MgCl_2$ was added to a 10 mM final concentration, and 3 µl of Hybridase Thermostable RNase H (Epicentre) was added, followed by a 30 min incubation at 45°C. The reaction was again purified by Zymo RNA Clean & Concentrator-5 column to deplete small RNA species, followed by treatment with DNaseI (Ambion) per manufacturer instructions and a final column clean-up to remove excess RNase H subtraction oligos.

100 ng of RNA was used for reverse transcription with 100 U TGIRT-III (InGex) for 2h at 57°C in the same TGIRT reaction conditions described above. We used 5-10 pmol of each gene-specific RT primer and successfully pooled up to six different RT primers in one reaction, using no more than 35 pmol total. DTT was prepared from powder directly before reverse transcription, and we omitted the denaturation step before reverse transcription due to low-level fragmentation of DMS-treated RNA at high temperatures. After moving the reaction to ice, 1 µl RNase H (Enzymatics, 5 U/µl) was added and RNA:DNA hybrids were degraded at 37°C for 20

min to release the cDNA. cDNA was purified using the ssDNA protocol for DNA Clean & Concentrator-5 columns (Zymo Research). We used the Advantage HF 2 PCR kit (Clontech) with high fidelity conditions for two-step PCR amplification, using 1/12 of the purified RT reaction and gene-specific primers targeting a single template with a target amplicon size of 300-600 nucleotides for low abundance RNA targets. High abundance RNAs, such as the yeast 18S rRNA, can be amplified in a single 1.8kb amplicon. Due to the high GC-content of the *FXR2* template, we used 200 mM NaCl instead of 75 mM KCl in the RT reaction buffer and the Advantage GC 2 PCR Kit (Clontech) for its amplification. The PCR program begins with 10 cycles at a 65°C annealing temperature to promote specificity, followed by 20-25 cycles at a 57°C annealing temperature. PCR bands were gel purified on a non-denaturing 8% TBE polyacrylamide gel (Invitrogen) and crushed, extracted, and EtOH precipitated as described above. NexteraXT (Illumina) was used to fragment and prepare amplicons (1ng) for sequencing. Tagmented amplicons were barcoded and amplified using 12 cycles of PCR, and barcoded libraries were cleaned using 1.5x (v/v) PCRClean beads (Aline Biosciences). Libraries were quantified using the Fragment Analyzer (Advanced Analytical) and subject to a final quantification by qPCR before sequencing by 50bp single-end reads on the HiSeq4000 (Illumina).

For the UMI-based RT-PCR, reverse transcriptase primers were designed with a random 10 nucleotide barcode, labeling each cDNA with a unique molecular index. Gene-specific variations of oMZ282 were used in the reverse transcription reaction described above, followed by Advantage HF 2 PCR with gene-specific variants of primers oMZ282 and oMZ283. Amplicons were purified by polyacrylamide gel and extracted as described above, and a second round of PCR was done with 20-25 cycles to add Illumina adaptors and indices for sequencing (oMZ284 and indexing primers). Libraries were constructed so the UMI was sequenced first using custom

Read1 sequencing primer oNTI202. We used the standard Illumina Read2 primer, and sequencing was done via MiSeq v2 2x150 (Illumina). See primer sequences in Table 3-S1.

Sequencing alignment and analysis. Fastq files were aligned and processed as outlined in Chapter 2. Due to empirically determined mutation enrichment from Nextera XT transposase insertion, we trimmed 7 nucleotides from the 5′ end of each read. Mismatches located within 3 nucleotides of an indel were also discarded for future analysis. The ratiometric DMS signal was calculated for each nucleotide as # mismatches / sequencing depth. Secondary structure models were visualized as described in Chapter 2.

Target-specific sequencing data prepared with NexteraXT was combined across both strand alignments, due to lack of strandedness after tagmentation. Transposase insertion is subject to primary sequence biases in transposase insertion, thus it is possible (although rare) to have amplicon regions that are poorly sampled and result in false positive bases with high ratiometric reactivity due to poor sequencing depth. After linker stripping with a length requirement for reads >100 nt from a 2x150 nt MiSeq run, target-specific sequencing data prepared with the UMI was collapsed to unique reads using FASTX-Collapser. Unique reads are, therefore, the combination of a unique molecular index and internal DMS-induced modifications, which add sequence diversity beyond the 10bp UMI.

Minimum average coverage calculation. Using 100 nt transcriptome windows, we chose the window with the highest total sequence coverage as representative coverage for the gene. We counted the fraction of genes from the hg19 RefSeq annotation that had an average mismatch coverage >120 mismatches at sequencing depths of 50, 100, and 200 million uniquely mapped reads. We extrapolated the data for 1 billion reads.

Computing the ROC curve for ribosomal RNA. This analysis was completed as previously described, using the yeast ribosome crystal structure[9] and the same considerations for solvent accessibility and removal of outliers by 90% Winsorization[6].

REFERENCES

1. Ben-Shem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334,** 1524–1529 (2011).
2. Aragón, T. *et al.* mRNA Targeting to ER Stress Signaling Sites. *Nature* **457,** 736–740 (2009).
3. Badis, G., Saveanu, C., Fromont-Racine, M. & Jacquier, A. Targeted mRNA degradation by deadenylation-independent decapping. *Mol. Cell* **15,** 5–15 (2004).
4. Hooks, K. B. & Griffiths-Jones, S. Conserved RNA structures in the non-canonical Hac1/Xbp1 intron. *RNA Biol.* **8,** 552–556 (2011).
5. Latrèche, L., Jean-Jean, O., Driscoll, D. M. & Chavatte, L. Novel structural determinants in human SECIS elements modulate the translational recoding of UGA as selenocysteine. *Nucleic Acids Res.* **37,** 5868–5880 (2009).
6. 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).
7. Chartrand, P., Meng, X. H., Singer, R. H. & Long, R. M. Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle in vivo. *Curr. Biol. CB* **9,** 333–336 (1999).
8. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10,** 623–629 (2013).
9. Ben-Shem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334,** 1524–1529 (2011).

CHAPTER FOUR

Novel RNA structure probing experiments *in vivo*, enabled by DMS-MaPseq

INTRODUCTION

A primary motivation for the development of DMS-MaPseq was its potential to enable the investigation of new RNA structure species *in vivo*. As described, the ability to use DMS-MaPseq for the targeted amplification of RNA greatly increases our ability to assess the folding status of species with low expression, and we sought to demonstrate this broad utility in several ways. First, we chose to trouble-shoot and demonstrate the use DMS for the *in vivo* chemical modification of RNA in an entirely new model system—the *Drosophila melanogaster* ovary, which also represents the first published usage of DMS for the RNA structure probing of a whole animal tissue. While application to *D. melanogaster* ovaries is a feature of the reactivity of dimethyl sulfate itself, as opposed to any particular component of the library preparation process for compatibility with next-generation sequencing, the developing oocyte is an excellent example of a system where many candidate RNAs are predicted to have structure-driven mechanisms that would be well served by the targeted DMS-MaPseq approach[1]. We also demonstrate the use of DMS-MaPseq to investigate the structure of the human *FXR2* mRNA. *FXR2* is expressed at low levels in human cells, meaning its RNA structure profile cannot be effectively captured in genome-wide experiments due to sequencing depth limitations, making it an excellent candidate for our targeted DMS-MaPseq approach. Additionally, *FXR2* was of particular interest to us given our previous discovery of its non-canonical translation initiation at a GUG start codon and exceptionally high GC content in its 5′ UTR. Using DMS-MaPseq, we were able to detect and functionally validate a regulatory RNA structure in the *FXR2* 5′ UTR region that affects expression of the protein.

Finally, in what we believe is truly the new frontier in RNA structure experimentation and discovery, we demonstrate the utility of DMS-MaPseq in separating RNA subpopulations to investigate their unique structure profiles. A given RNA species could adopt many heterogenous

RNA structure conformations depending on its specific biological state. Broadly, it is not known how RNA structure varies based on the proteins that are bound, the location of the RNA in the cell, its engagement in processing or degradation, or unique non-quantitative sequence features like endogenous modifications or single nucleotide polymorphisms. In this chapter, we demonstrate the utility of our DMS-MaPseq approach to distinguish the RNA structure profiles between human alleles and to independently assess the structural differences in yeast pre-mRNA relative to its mature processed counterparts.

RESULTS

DMS-MaPseq for *D. melanogaster* ovaries

Drosophila melanogaster has served as a premier system for studying mRNA localization and translational control during development because dramatic developmental changes occur in the absence of transcription and mRNA degradation. The future embryonic body axes are established prior to fertilization by localization of a large number of mRNAs during oogenesis[1]. The *cis*-signals directing mRNAs to different poles of the oocyte are poorly understood but have been shown in some cases to involve RNA structure[2–4]. Here, we present targeted DMS-MaPseq data from the *in vivo* structure probing of *D. melanogaster* ovaries, which yields excellent structure data consistent with the *oskar* and *gurken* mRNA structures responsible for localization[4,5] (Fig. 4-1a, Fig. 4-S1). These data also represent the first example of RNA structure probing in an animal tissue. Thus, DMS-MaPseq overcomes a key experimental challenge for understanding the role of RNA structure during oogenesis.

A highly structured region influences non-canonical translation initiation of the low-expression *FXR2* mRNA

We recently discovered that translation of the mammalian *FXR2* (Fragile X Mental Retardation, Autosomal Homolog 2) gene initiates predominantly at a GUG codon significantly upstream of the previously annotated AUG initiation site[6]. Due to the extreme GC content (80%) of the first exon of *FXR2*, which encodes its 5′ UTR and early CDS, we hypothesized that a stable RNA structure may contribute to the mechanism of GUG-initiation. We used DMS-MaPseq *in vitro* data to develop a secondary structure model with RNAfold informed by experimental constraints[7]. This revealed two highly stable putative structures flanking the GUG intiation codon (Fig. 4-1b, Fig. 4-S2 a,b; free energy < -31 kcal/mol). To explore the functional consequences of *FXR2* translation initiation, we mutated these putative structures in a reporter construct comprised of the *FXR2* exon1 sequence fused to eGFP and observed a drop in protein levels upon mutating either structure (Fig. 4-S2 c,d). Importantly, compensatory mutations to restore the predicted RNA structures also restored eGFP levels, implicating the structure itself as a functional modulator of translation initiation for *FXR2*.

Structure probing of RNA species in multiple conformations

In the complex environment of the cell, the structure of an RNA molecule is likely to vary based on the biological process in which it is engaged, such as maturation, translation, protein binding, and degradation. To date, *in vivo* RNA structure probing techniques reliant on RT truncation necessitate the assessment of RNA structure signal across an ensemble average population of RNA molecules, thereby blurring signal from any structural heterogeneity that might exist. In the case of structural heterogeneity derived from a ribosnitch, i.e. a single nucleotide polymorphism that yields a local RNA structure rearrangement, the interpretation of *in vitro* RNA folding status differs greatly when DMS-MaPseq data from the two human *MRPS21* ribosnitch alleles[8] are analyzed together or separately. Allele-specific analysis of the data reveals two distinct and mutually exclusive structures, which are not detectable from the combined allele analysis (Fig. 4-2a).

With the advancement of DMS-MaPseq, it is possible to investigate *in vivo* RNA structure heterogeneity, utilizing either the co-occurrence of multiple modifications per molecule or the specific amplification of RNA subpopulations. Of particular interest are isoform-specific RNA structures that have been proposed in pre-mRNAs versus their mature translated counterparts, such as RNA structures which influence splice site selection[9] or require unwinding by the ribosome, influencing translation initiation and processivity[10,11]. To investigate the specific RNA structure profiles of yeast pre- and mature mRNAs, we used intron- or exon-specific RT primers to separately amplify each isoform of two yeast ribosomal protein genes using the targeted DMS-MaPseq approach. Comparison of RNA structure signal in the common exon1 sequence between the *RPL14A* and *RPL31B* pre-mRNAs and their respective mature amplicons reveals surprisingly little structure difference between isoforms (Fig. 4-2b and Fig. 4-S3, respectively). These ribosomal protein genes are highly translated, but their common exon1 sequence appears similarly structured in both the mRNA and untranslated pre-mRNA, suggesting that local RNA structure rapidly refolds after translation. While we focus here on a limited number of messages, this approach will broadly enable the analysis of different RNA isoforms, including lowly expressed species.

FIGURES

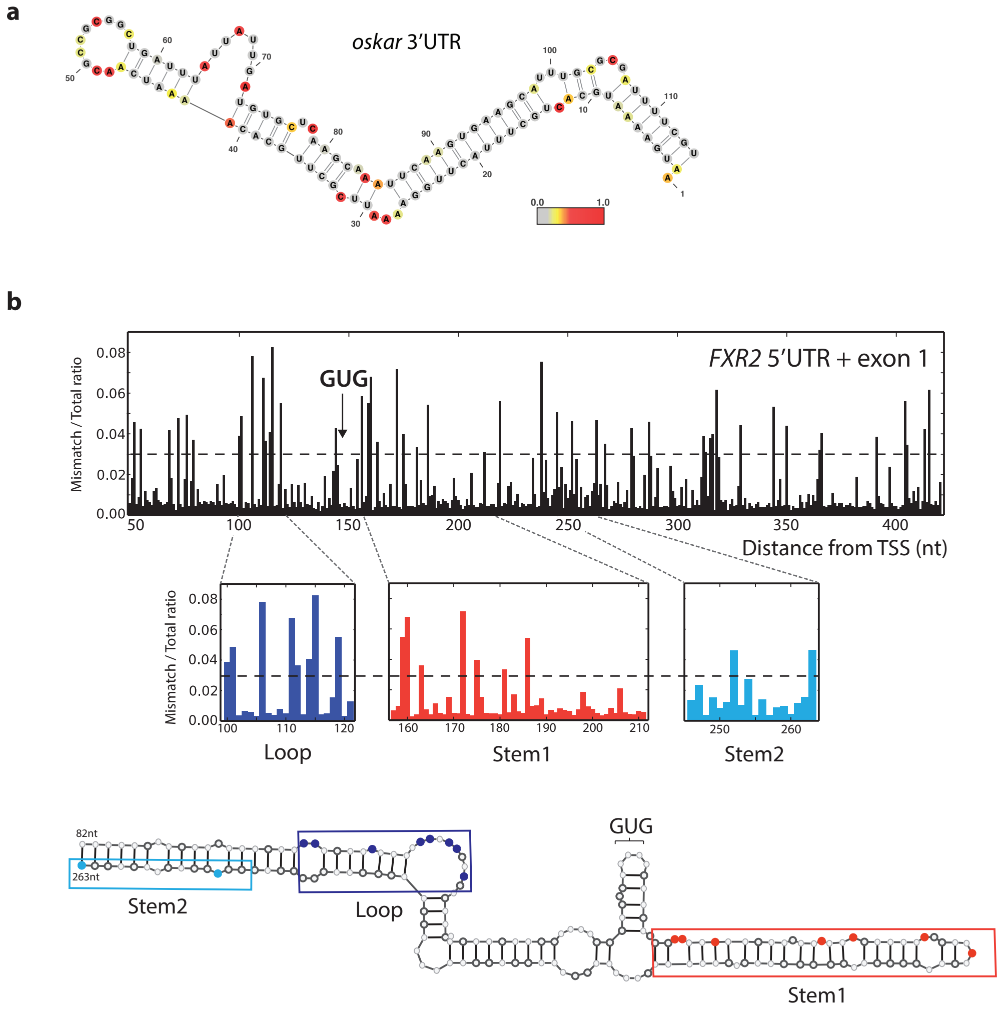


Figure 4-1 | Novel experimental applications for *in vivo* RNA structure probing. **a,** *oskar* 3′ UTR mRNA positive control structure from target-specific priming with nucleotides colored by *in vivo* DMS reactivity in *D. melanogaster* ovaries. DMS reactivity calculated as the ratiometric DMS signal per position normalized to the highest number of reads in displayed region, which is set to 1.0. **b,** Ratiometric DMS-MaPseq from targeted amplification of the human *FXR2* 5′ UTR and exon1 sequence. Nucleotides accessible to DMS are noted with a value >0.03. Position 1 corresponds to chromosome XVII:7614897. Secondary structure modeling reveals two stable stems flanking the GUG initiation site.

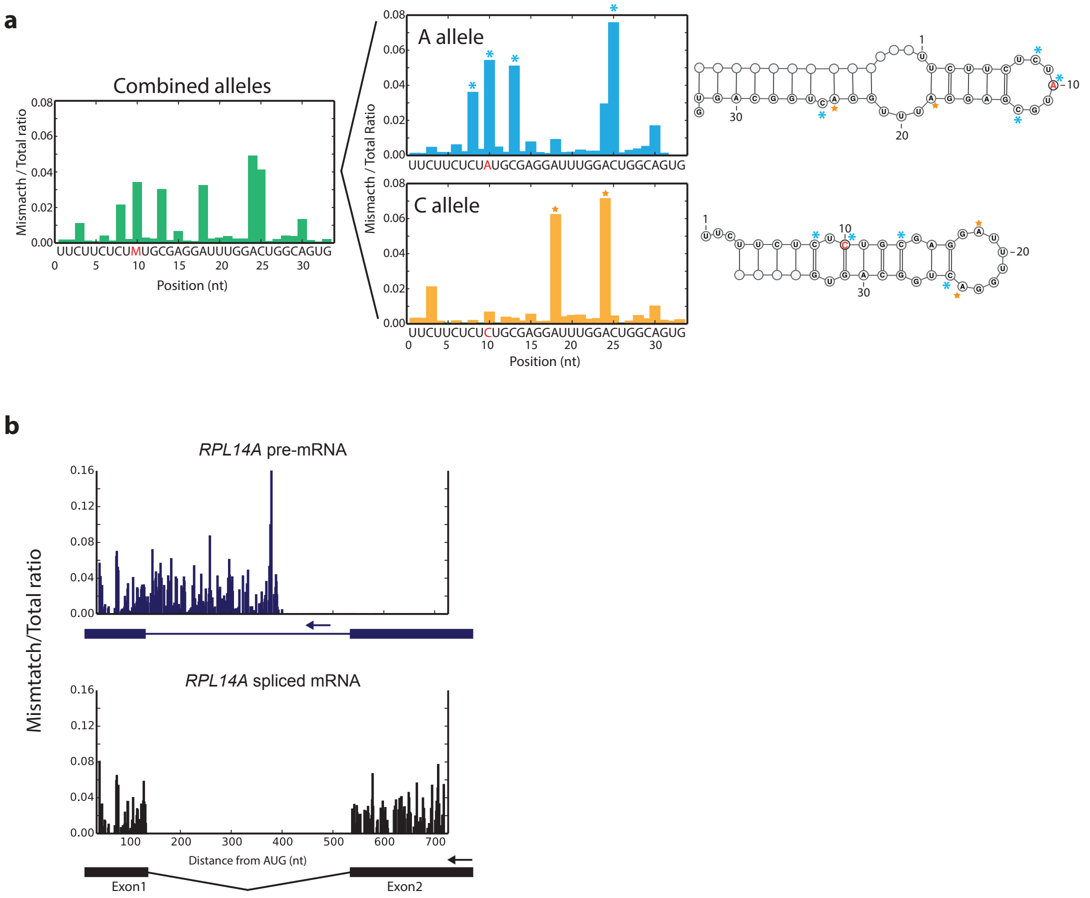


Figure 4-2 | Investigating RNA structure heterogeneity with DMS-MaPseq. **a,** Regions of heterogeneous structure exhibit indistinguishable structure signals when combined but can be distinguished by DMS-MaPseq, illustrated by ratiometric DMS-MaPseq data derived from the human *MRPS21* ribosnitch A/C alleles.
b, Targeted DMS-MaPseq data specific for the yeast *RPL14A* pre-mRNA and spliced mRNA isoforms reveal minimal structure difference in the common exon1 sequence (*r* = 0.88). Ratiometric *in vivo* DMS-MaPseq data is plotted with isoform-specific RT primer locations noted with arrows.

SUPPLEMENTAL INFORMATION

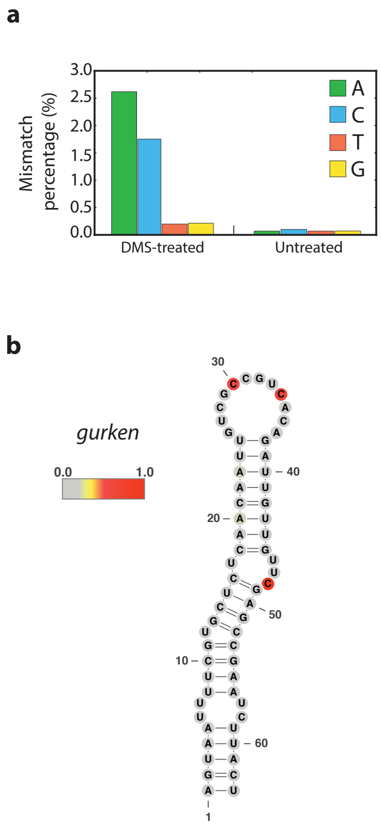
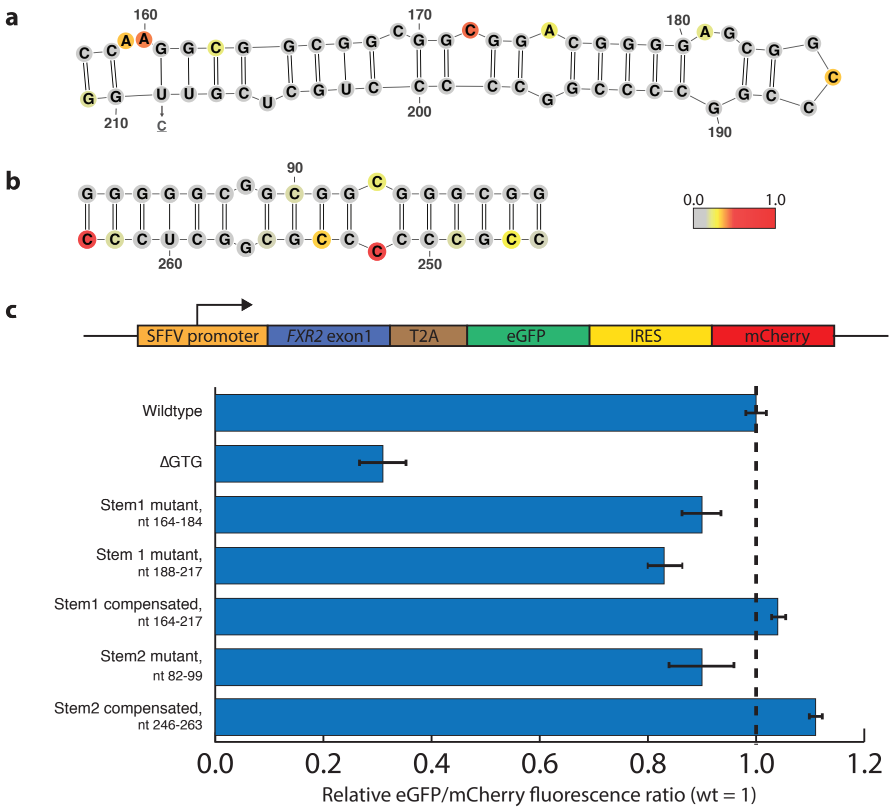


Figure 4-S1 | Targeted DMS-MaPseq in *D. melanogaster* ovaries. **a,** Total mismatch percentage observed on each nucleotide from *in vivo* DMS-treated and untreated *D. melanogaster* RNA. **b,** *gurken* mRNA positive control structure with nucleotides colored by *in vivo* DMS reactivity. DMS reactivity calculated as the ratiometric DMS signal per position normalized to the highest number of reads in displayed region, which is set to 1.0.

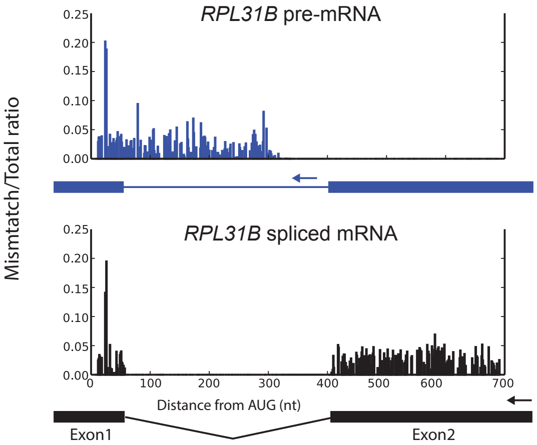


d

Mutation Type	Mutation Sequence
Stem1 mutant 164-184	TTGCCATCcCCaGCCCCagGC
Stem1 mutant 188-217	GCgGGGGCgCGGCGGCGtGCGG
Stem1 compensated 164-217	TTGCCATCcCCaGCCCCagGCgccGCgGGGGCgCGGCGGCGtGCGG
Stem2 mutant 82-99	CCCTCGgCGCCcCCCGCC
Stem2 compensated 246-263	GGCGGGcGGCGCCGGGG

Figure 4-S2 | Fluorescent reporter constructs with RNA structure mutations confirm function of a highly stable structure in *FXR2* translation. **a, b,** Predicted stems 1 and 2 in the human *FXR2* 5′ UTR and first exon, with nucleotides colored by DMS reactivity. DMS reactivity calculated as the ratiometric DMS signal normalized to the highest reactive base. **c,** *Top, FXR2* reporter construct design. The 5′ UTR and first exon of human *FXR2* ΔATG is fused to a T2A and in-frame eGFP lacking its initial AUG, such that mutations to the coding region of *FXR2* will not affect stability of the eGFP protein. To internally control for transfection and transcription efficiency, mCherry driven by an internal ribosome entry site was included downstream. *Bottom,* fluorescence measurements following transient transfection of *FXR2* reporter constructs into HEK 293T cells. The eGFP/mCherry ratio was calculated for transfection replications of each construct and scaled relative to the wildtype construct, which was set to 1.0. Error bars represent one standard deviation. This analysis reveals a drop in eGFP levels upon mutating the predicted *FXR2* structure and a full recovery of eGFP levels after compensatory mutation. Basal levels of protein expression in ΔGTG mutant likely reflects translation initiation at other NUG codons. **d,** Sequences for mutations assayed in *FXR2* reporter system, nucleotides predicted to be unpaired are shown in lowercase typeface and were not mutated.

a



b

Gene	*r* value	
	pre-mRNA replicates	pre- vs spliced
RPL14A	0.96	0.88
RPL31B	0.98	0.95

Figure 4-S3 | RNA structure does not vary between the pre-mRNA and spliced mRNA isoforms of yeast ribosomal protein genes. **a,** Targeted DMS-MaPseq data specific for the yeast *RPL31B* pre-mRNA and spliced mRNA isoforms reveal minimal structure difference in the common exon1 sequence. Ratiometric DMS-MaPseq data is plotted with isoform-specific RT primer locations noted with arrows. **b,** Exon1 DMS-MaPseq structure signal correlation (Pearson's *r* value) across pre-mRNA and spliced mRNA isoforms and between isoform-specific replicates.

Table 4-S1.

Primers used in this chapter.

name	purpose	sequence (5' to 3')
oMZ289	*RPL14A* RT/Rev PCR primer, intron	ACGATGGAAGGCATGGTTTAATATTTGAGGAAACATGG
oMZ290	*RPL14A* RT/Rev PCR primer, exon	AGCCTTAGCCAAAGCCTTCTTGACAGTGTA
oMZ291	*RPL14A* Fwd PCR primer	TGTCCACCGATTCTATTGTCAAGGCTTCTAACTGG
oMZ292	*RPL31B* RT/Rev PCR primer, intron	AAGGTGGAACTAAAGCATCACGCCAAAAACATCG
oMZ293	*RPL31B* RT/Rev PCR primer, exon	CGTACCCCGAAAGCAGCTCTGTTTGTGTAAT
oMZ294	*RPL31B* Fwd PCR primer	TGCACGAGCAGATAATCCAAAGTACTTGAAAATGGCC

Table 4-S2.

Plasmids used in this chapter.

name	description
pJW1643	pLeGO-ic2 pSFFV-FXR2exon1(wt)-3xFLAG-TEV-HA-T2A-eGFP-IRES-mCherry
pJW1644	pLeGO-ic2 pSFFV-FXR2exon1(ΔGTG)-3xFLAG-TEV-HA-T2A-eGFP-IRES-mCherry
pJW1645	pLeGO-ic2 pSFFV-FXR2exon1(mut, 164-184)-3xFLAG-TEV-HA-T2A-eGFP-IRES-mCherry
pJW1646	pLeGO-ic2 pSFFV-FXR2exon1(mut, 188-217)-3xFLAG-TEV-HA-T2A-eGFP-IRES-mCherry
pJW1647	pLeGO-ic2 pSFFV-FXR2exon1(comp, 164-217)-3xFLAG-TEV-HA-T2A-eGFP-IRES-mCherry
pJW1648	pLeGO-ic2 pSFFV-FXR2exon1(mut, 82-99)-3xFLAG-TEV-HA-T2A-eGFP-IRES-mCherry
pJW1649	pLeGO-ic2 pSFFV-FXR2exon1(comp, 246-263)-3xFLAG-TEV-HA-T2A-eGFP-IRES-mCherry

MATERIALS AND METHODS

Dimethyl sulfate (DMS) modification. For *D. melanogaster* oocytes, we dissected ovaries from ~100 flies (OreR strain) in 250 µl 1X PBS. We added 250 µl DMS for 5 min at 26°C with shaking at 500 rpm. To stop the reaction, we added 1 ml of 30% BME (v/v) and transferred the oocytes to a sieve, where they were washed three times in 30% BME and two times with sterile water. Finally, the ovaries were collected and re-suspended in 1mL of Trizol and 10 µl BME, and total RNA was extracted.

Library generation, targeted DMS-MaPseq. Targeted amplification of RNA was completed as described in Chapter 3. Primers used are listed in Table 4-S1.

Cloning and transfection experiments. The plasmid construct in Figure 4-S2 was derived from the ∆ATG *FXR2*exon1-eGFP-IRES-mCherry plasmid described in *Fields et al.*[6]. A gBlock (IDT) was ordered containing a 43bp FXR2-3xFLAG-T2A-AgeI-40bp eGFP fragment for HiFi assembly (NEB) into the linearized plasmid backbone. This wildtype plasmid was used as the PCR template for *FXR2* mutations, which were designed as overhangs on primers against the relevant portion of the *FXR2* exon1 sequence, resulting in 5′ and 3′ fragments with overlapping mutated regions for HiFi assembly into the linearized wildtype backbone. Successful amplification of fragments was confirmed by running a fraction on an agarose gel and the remainder purified using DNA Clean & Concentrator-5 columns (Zymo) or, in the case of contaminating PCR bands, purified via agarose gel and MinElute gel extraction (Qiagen). Common cloning primers for *FXR2* amplification from the plasmid are 5′-CTCACTCGGCGCGCCAGTC-3′ (5′ *FXR2* fragment, forward) and 5′-TATAGTCCCCGTCGTGATCCTTGTA-3′ (3′ *FXR2* fragment, reverse). Inserts in all analyzed constructs were confirmed by Sanger sequencing (Molecular Cloning Laboratories). Plasmids are listed in Table 4-S2.

For fluorescence measurements, HEK 293T cells were grown as described and transfected with plasmids using TransIT-LT1 (Mirus) two days prior to data collection. eGFP and mCherry fluorescence were quantified using an LSR-II flow cytometer (BD Biosciences). Two plasmids for each type of mutation were assayed for fluorescence, serving as biological duplicates.

Ribosnitch RNA preparation. dsDNA corresponding to the human *MRPS21* sequences shown below were *in vitro* transcribed, mixed, and folded by denaturing at 95°C followed by a brief incubation at 37°C in 350 mM sodium cacodylate buffer and 6 mM $MgCl_2$. 10% DMS (v/v) was added, and the sample was incubated for 10 min at 37°C. The reaction was stopped by placing on ice and adding BME to 30% final volume. The RNA was then purified by RNA Clean & Concentrator-5 column (Zymo), and the small RNA fraction was collected and prepared for sequencing as described in the genome-wide strategy above.

MRPS21 A allele, 5′-TGCTGCCATCTCTTTTCTTCTCTATGCGAGGATTTGGACTGGCAGTG-3

MRPS21 C allele, 5′-ATCTCTTTTCTTCTCTCTGCGAGGATTTGGACTGGCAGTGAGAATAAGAGACAA-3′

REFERENCES

1. Jambor, H. *et al.* Systematic imaging reveals features and changing localization of mRNAs in Drosophila development. *eLife* **4,** (2015).
2. MacDonald, P. M. bicoid mRNA localization signal: phylogenetic conservation of function and RNA secondary structure. *Dev. Camb. Engl.* **110,** 161–171 (1990).
3. Bullock, S. L., Ringel, I., Ish-Horowicz, D. & Lukavsky, P. J. A′-form RNA helices drive microtubule-based mRNA transport in Drosophila. *Nat. Struct. Mol. Biol.* **17,** 703–709 (2010).
4. Jambor, H., Brunel, C. & Ephrussi, A. Dimerization of oskar 3' UTRs promotes hitchhiking for RNA localization in the Drosophila oocyte. *RNA N. Y. N* **17,** 2049–2057 (2011).
5. Van De Bor, V., Hartswood, E., Jones, C., Finnegan, D. & Davis, I. gurken and the I factor retrotransposon RNAs share common localization signals and machinery. *Dev. Cell* **9,** 51–62 (2005).
6. Fields, A. P. *et al.* A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol. Cell* **60,** 816–827 (2015).
7. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol. AMB* **6,** 26 (2011).
8. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505,** 706–709 (2014).
9. Meyer, M., Plass, M., Pérez-Valle, J., Eyras, E. & Vilardell, J. Deciphering 3'ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. *Mol. Cell* **43,** 1033–1039 (2011).
10. Babendure, J. R., Babendure, J. L., Ding, J.-H. & Tsien, R. Y. Control of mammalian translation by mRNA structure near caps. *RNA* **12,** 851–861 (2006).
11. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324,** 255–258 (2009).

CHAPTER FIVE

Discussion and future perspectives

SUMMARY

Here we establish DMS-MaPseq as a robust and simple tool that, in many respects, serves as the premier technique for the quantitative analysis of RNA secondary structure *in vivo* by improving the inherent quality of the structure data, enabling qualitatively new types of structure to be gathered, and greatly expanding the repertoire of RNAs that can be analyzed. Three key features of DMS-MaPseq enable these improvements. First, the minimization of library biases and the high signal-to-noise ratio in DMS-MaPseq data yield an inherently ratiometric and quantitative readout of structure with single nucleotide resolution. Second, the ability to use mutational profiling *in vivo* enables experiments that can provide key insights into RNA structure heterogeneity in the complex cellular environment—an important biological question that remains poorly understood. In addition to the selective amplification of isoforms to investigate structure differences as we demonstrate here, further applications of DMS-MaPseq include *in vivo* single-molecule analyses of the co-occurrence of DMS modifications to identify RNA structure subpopulations empirically.

Finally, perhaps the most practical advance provided by DMS-MaPseq is the ability to selectively amplify RNA targets, which drastically expands the range of RNA species suitable for *in vivo* RNA structure probing and does so with low sequencing costs and simple experimental implementation. Together, these advances enable a wide range of future studies, including a comprehensive investigation of RNA structure differences between pre-mRNA and mature mRNA isoforms or between alternatively spliced mRNA variants. Isoform-specific experiments may also address whether specific RNA structures function in splicing or other pre-RNA processing steps, such as those unique to pre-rRNA processing, and how translation influences mRNA folding status relative to its untranslated, immature state. In theory, our *in vivo* mutational profiling approach could also be applied to chemical modification by SHAPE, but the modification introduced by the SHAPE variant with well-validated *in vivo* reactivity, NAI-N3[1], is bulky and presents a challenge in finding an RT enzyme capable of reading through it.

In the future, DMS-MaPseq could also be combined in conjunction with other experimental techniques used for the genome-wide discovery of endogenous mRNA modifications, including the sequencing-based mapping of pseudouridines or sites of m^6A methylation[2–5]. These endogenous modifications occur on only a subset of their RNA targets, imparting interesting questions about the influence of a natural modification on local structure and its functional consequences. Combined with the single-molecule aspects of DMS-MaPseq, it would be possible to analyze endogenously modified RNA subpopulations in a single experiment to address these questions. It is the versatility of DMS-MaPseq that makes it a transformative tool for *in vivo* RNA structure probing, allowing for more comprehensive investigations into the biological relevance of RNA structures than ever before.

The future applications of DMS-MaPseq are numerous, and we are particularly interested in using the technique to empirically resolve *in vivo* RNA structure heterogeneity based on the clustering of DMS-MaPseq reactivity profiles alone. RNA is engaged in many different biological processes throughout its lifetime, which likely impose different requirements on the local RNA structure context. After transcription, an RNA may be capped, spliced, exported to the cytoplasm, modified, translated, localized, sequestered, and degraded. An RNA may also function as a scaffold or an enzyme during its lifetime. Depending on their biological state, it is known that RNA structure can vary[6,7]. However, due to complete sequence identity, we have been unable to experimentally distinguish these types of RNA subpopulations from one another. With the ability to collect multiple pieces of RNA structure information within a single sequencing read, the increased information content enables computational approaches to cluster based on structure classes[8]. Current DMS treatment conditions yield ~1 DMS modification per 50 nucleotides, which does not produce a robust information content for clustering or co-occurrence algorithms. Thus, additional experiments must be done to identify the ideal DMS modification threshold. Specifically, DMS treatment conditions should maximize modifications to a point at which the high number of modifications does not alter the RNA structure itself. For example, a dynamic or breathing RNA structure might be forced into an unfolded state if DMS modifications block natural refolding. Additionally, as the DMS modification rate increases, troubleshooting the reverse transcription conditions may also be necessary due to a suspected propensity for the TGIRT enzyme to pause when decoding modifications, potentially resulting the loss of signal on the 5' end of the RNA fragment in this way. An additional and exciting application of DMS-MaPseq is its adaption for single-molecule RNA structure experiments. A long-read sequencing technology, such as SMRT Sequencing from Pacific Biosciences, combined with DMS-MaPseq and a high DMS modification rate could transform our ability to infer RNA structure empirically.

In summary, methods for experimentally determining *in vivo* RNA structure have greatly improved in recent years. Chemical-based probing techniques are now more quantitative and can be used for genome-wide or targeted applications. Target-specific amplification broadly democratizes RNA structure probing experiments, boasting an easy technical implementation and low sequencing demands that make it an attainable experiment for smaller labs. We can now experimentally assess RNA structure variation in the case of small sequence differences that can be used to either experimentally or computationally distinguish between species. With these versatile tools for experimentally determining RNA structure *in vivo*, more RNA structures can be identified, their functions investigated, and the role of RNA folding in cellular mechanism will be better understood.

REFERENCES

1. Spitale, R. C. *et al.* Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519,** 486–490 (2015).
2. Carlile, T. M. *et al.* Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* **515,** 143–146 (2014).
3. Schwartz, S. *et al.* Transcriptome-wide mapping reveals widespread dynamic regulated pseudouridylation of ncRNA and mRNA. *Cell* **159,** 148–162 (2014).
4. Meyer, K. D. *et al.* Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3' UTRs and Near Stop Codons. *Cell* **149,** 1635–1646 (2012).
5. Dominissini, D. *et al.* Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485,** 201–206 (2012).
6. Dethoff, E. A., Chugh, J., Mustoe, A. M. & Al-Hashimi, H. M. Functional complexity and regulation through RNA dynamics. *Nature* **482,** 322–330 (2012).
7. Montange, R. K. & Batey, R. T. Riboswitches: emerging themes in RNA structure and function. *Annu. Rev. Biophys.* **37,** 117–133 (2008).
8. Homan, P. J. *et al.* Single-molecule correlated chemical probing of RNA. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 13858–13863 (2014).

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature	8/15/2016 Date