Single molecule structure sequencing reveals RNA structural dependencies, breathing and ensembles

Teshome Tilahun Bizuayehu[1,2], Kornel Labun[1], Kirill Jefimov[1], Eivind Valen[1,2*]

Affiliations

[1] Computational Biology Unit, Department of Informatics, University of Bergen, Norway.

[2] Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen 5008, Norway

*To whom correspondence should be addressed: eivind.valen@gmail.com

ABSTRACT

RNA molecules can form secondary and tertiary structures that determine their localization and function[1]. Using enzymatic or chemical probing together with high-throughput sequencing, secondary structure can be mapped across the entire transcriptome[2]. A limiting factor, however, is that only population averages can be obtained, since each read is an independent measurement. In addition, no information about structural heterogeneity across molecules or dependencies within each molecule is accessible. Here, we present Single Molecule Structure sequencing (SMS-seq) that combines structural probing with native RNA sequencing to provide non-amplified, structural profiles of individual molecules. Each RNA is probed at numerous bases enabling the discovery of dependencies and heterogeneity of structural features. In addition to revealing known structural features of mRNAs, SMS-seq shows compartmentalization of structural dependencies across CDSs and 3'UTRs. Finally, we show that SMS-seq can capture structural breathing, tertiary interactions and dynamics of riboswitch ligand binding.

RESULTS AND DISCUSSION

High-throughput RNA structure probing commonly use reagents that react with nucleotides to interrogate the chemical accessibility of a region. The position of chemical modification is then detected by reverse transcription drop-off and/or mutation [2]. Among RNA structure probing reagents, diethyl pyrocarbonate (DEPC) has been used to map structures of RNAs *in vitro* [3–5]. DEPC carbethoxylates unpaired adenosine at N-7 position [6] (Fig. 1A) – a modification which is detectable on the Oxford Nanopore MinION sequencer (Supp. Fig. 1). To assess how well this approach could capture structural information, we synthesized and probed a hairpin RNA with a highly predictable structure (Supp. Fig. 2). We sequenced direct RNA treated with DEPC and compared these with control samples (Fig. 1B) obtaining 3093 and 4687 full-length reads, respectively. To detect modified bases, we contrasted the signal from each base of each read with the control sample using modifications called by Tombo[7] at a threshold of 0.2 (Fig. 1D) corresponding to a false discovery rate of 16% across untreated regions. This gave an average of 7.43 modified bases in the hairpin over full-length reads (Supp. Fig. 3). DEPC-induced modification sites were detectable and obtained near-perfect consensus classification for open versus single classification error in stem regions, which could be as result of conformational heterogeneity (Fig. 1C, E, Supp. Fig. 4). More interestingly, individual bases on individual reads achieved a sensitivity of 0.50 and specificity of 0.74 (Fig. 1E). However, since most accessible bases are not modified in every read, the number of false negatives is likely inflated and the sensitivity is underestimated. False positives in the inaccessible regions come, at least partially, due to the nature of nanopore sequencing. Some k-mers are more sensitive than others to be detected as modified and where neighboring, potentially modified, bases can influence the

current signal. However, some modifications within stems are likely to represent real, low-frequency conformations where these bases are transiently exposed and modified (RNA breathing)[8]. To explore this, we folded individual reads with ViennaRNA Package [9] using hard constraints prohibiting modified bases from being part of a stem. Reassuringly, the most abundant structure was the expected structure indicating that 45.1% of reads have no modification conflicting with the hairpin structure. These were however heavily modified in the expected open regions, showing that the hairpin structure is indeed the most favourable conformation. Most alternative structures are only from a single read, and potentially represent noise in calling of modifications and/or transient structures (Fig. 1F). However, notably, more frequent structures showed individual bases modified in the hairpin stem and/or more open loop regions (Fig. 1G). This result indicates that SMS-seq can measure conformation heterogeneity and RNA breathing.

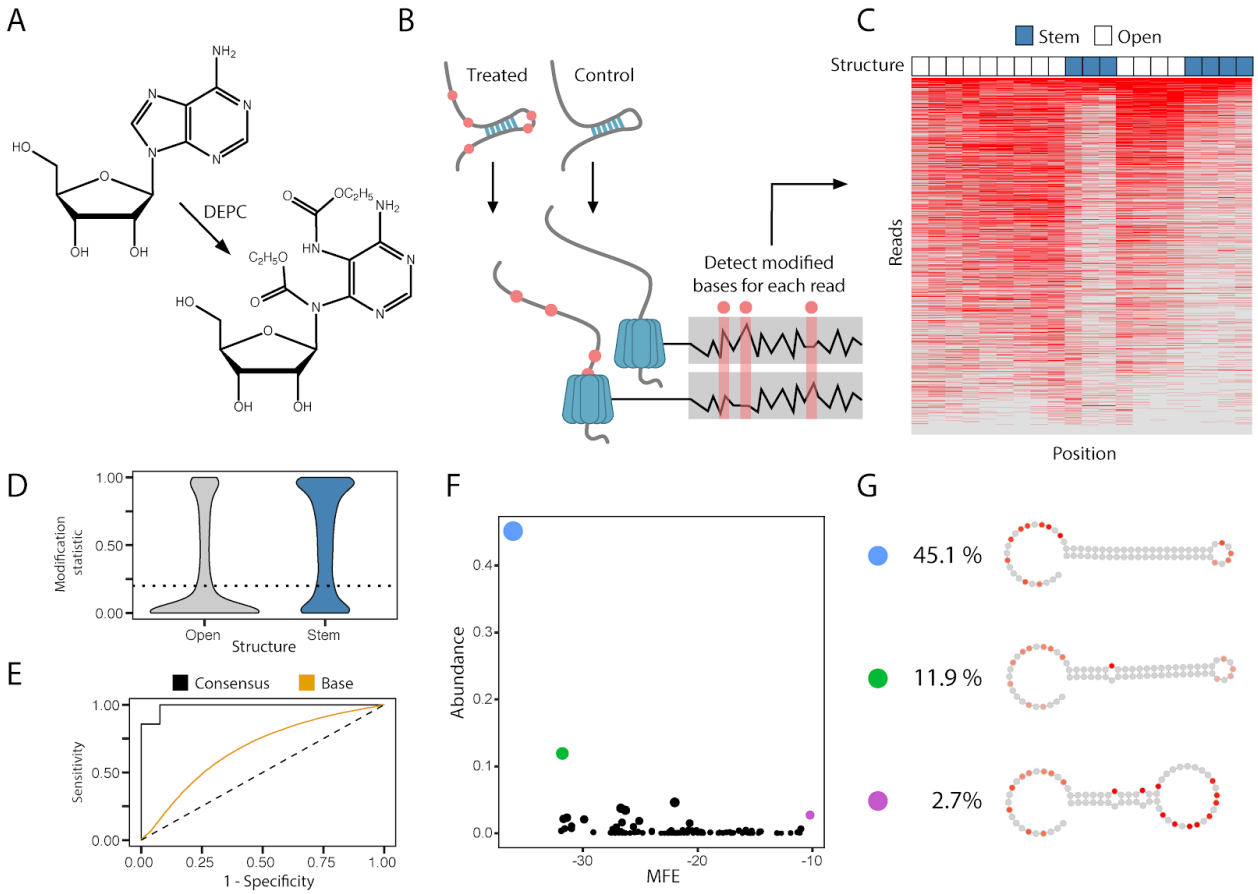

Fig 1. Oxford nanopore direct RNA structure profiling of a hairpin RNA.

A) Molecular structure of Adenosine modification by DEPC that modifies structurally unconstrained adenosine by opening the imidazole ring. **B**) SMS-seq workflow. Structures are probed with DEPC and the current signal of each nucleotide that passes through the pore is recorded. Modified sites (red) are detected by comparing the current signal from treated samples to the control samples. **C**) Heatmap showing 3093 full-length reads (y-axis) with modified (red) and unmodified (gray) nucleotides at each position (x-axis) for each read. Reads are ordered by overall modification rate. The legend above (Structure) shows which nucleotides are accessible (white) and inaccessible (blue) in the hairpin structure. **D**) Tombo modification statistics for adenosine bases. Bases are called modified (open) when their statistic is below 0.2 (dotted line) corresponding to an FDR of 16%. **E**) ROC curve for the hairpin at the consensus level (black) and individual bases (orange). Random model shown with dotted line. **F**) Groups of

reads folding to the same structure shown by abundance (y axis) and minimum folding energy (x axis). **G**) Selected structures with modified bases colored (red circles).

Methods based on enzymatic cleavage or termination during reverse transcription can only capture independent measurements from a single location. Unlike these, SMS-seq can modify and record numerous exposed bases per molecule. SMS-seq is therefore uniquely capable of capturing dependencies of open regions across individual molecules. We first assessed the extent to which we could obtain information on the dependencies in the hairpin, by calculating the adjusted mutual information (AMI) between all positions (Methods). Since calling modifications in nanopore sequencing is influenced by neighboring bases, we normalized the AMI by the expected AMI across all bases spaced at the same distance in data prepared from yeast mRNA (Supp. Fig. 5, Methods) giving a distance-normalized AMI (dAMI). This revealed that bases that are part of the same structural unit share a high dAMI. In particular, the two co-dependent strands in the stem of the hairpin had a modification state that depend on each other (Fig. 2A). While the dAMI can infer dependencies, it relies on there being variation in the modification state at both positions. For positions that are predominantly in one state, the dAMI can therefore be weak. We therefore calculated to what extent bases are co-observed in the same state. Specifically, we calculated co-openness, the probability that one base is modified given that the other is modified; and co-agreement, the fraction of reads that they share the same state (Fig. 2B). These metrics highlighted the open and stem regions, respectively. Together this demonstrated that SMS-seq can capture structural co-dependencies in stable structures.

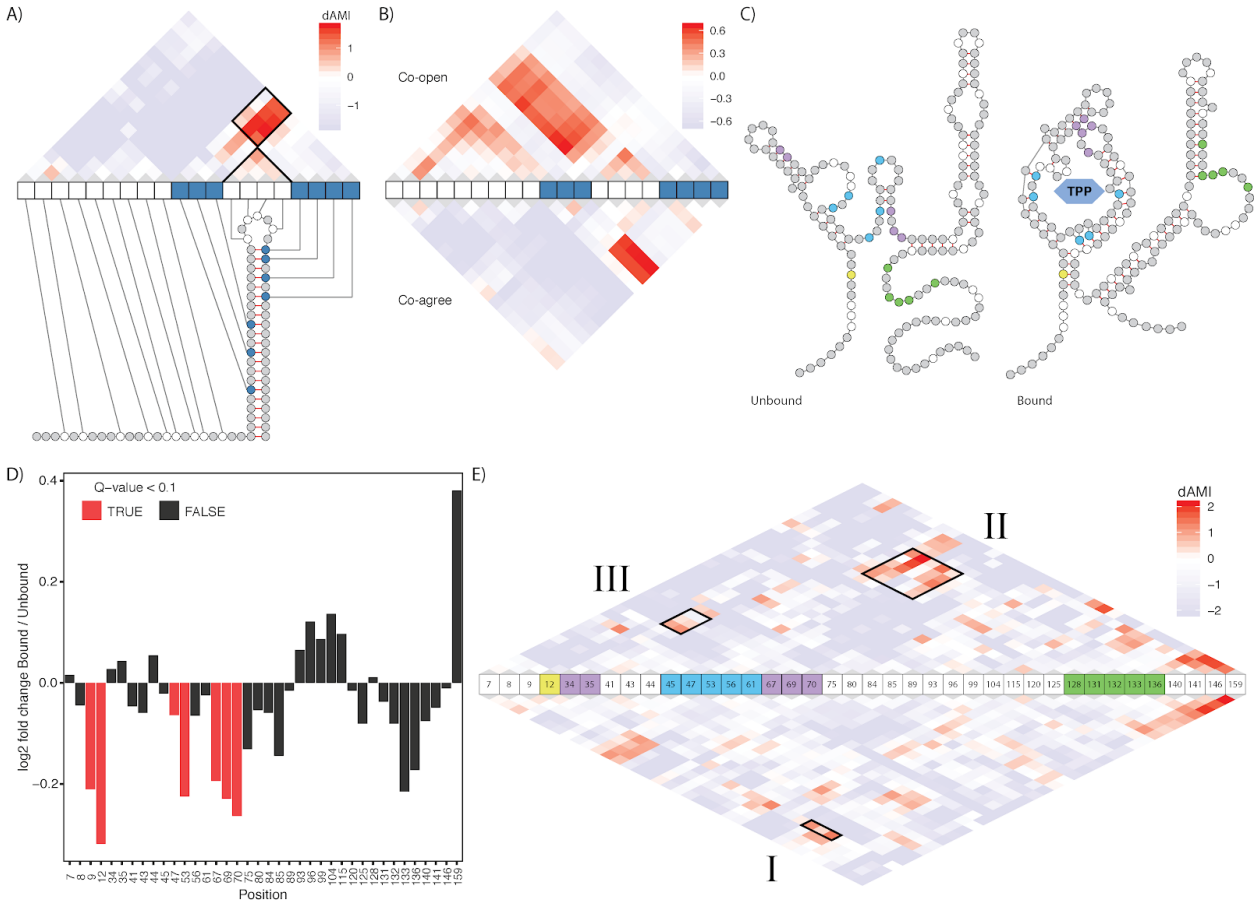

Fig 2. SMS-seq reveals structural dependencies in RNA

A) dAMI dependencies for pairs of bases in the synthetic hairpin construct. **B)** Normalized co-openness and co-agreement scores for the synthetic hairpin construct. **C)** Predicted secondary structure obtained from previous study[10] for the TPP riboswitch in unbound (left) and bound (right) configuration. Domains are indicated in colors: first stem (yellow), aptamer domain (blue), tertiary interaction of loop-stem (purple), ribosome binding site (green). Only adenosine residues are coloured. **D)** Log2 fold change in accessibility between bound and unbound states of TPP. Only adenosine residues are depicted. Red bars indicate significant changes ($q < 0.1$) in ligand-binding domain. **E)** dAMI dependencies between pairs of bases for the bound (upper) and unbound (lower) state of the TPP riboswitch. Domains are colored according to subfigure C.

We next asked whether SMS-seq could capture the dynamics and dependencies of structured

riboswitches. We probed two riboswitches in their ligand bound and unbound state: thiamine pyrophosphate (TPP[10], Fig. 2C) and *F. nucleatum* Flavin mononucleotide (FMN[11], Supp. Fig. 6A). Both bound and unbound TPP structures had similar modifications profiles (Supp. Fig. 7) indicating preformation of structure in the absence of the ligand. Upon binding, both riboswitches became overall less accessible, as expected (Fig 2D, Supp. Fig. 6B). Specifically, for the TPP riboswitch, nucleotides in the aptamer domain (positions 11–88) as well as those that encompass the ribosome binding site (position 125-146) are less accessible in the bound state. This is consistent with reduced access for ribosomes and translational inhibition as previously observed [12]. Our dAMI analysis further revealed that SMS-seq captures the dependency of the first stem (position 12) on the overall architecture of TPP (Fig. 2E, diagonal line extending from 12) and in particular the ribosome binding site (Fig. 2E, box I). The dependency of the expression domain on the conformational change of TPP binding pocket in the presence of TPP is also evident (Fig. 2E, box II). Interestingly, dAMI also revealed that SMS-seq is capable of capturing the tertiary interactions between the two helices (Fig. 2E, box III), as previously reported through footprinting experiments and X-ray crystallography [10]. Similar observations can be made for the FMN where SMS-seq captures the extensive interactions (stem-loop and loop-loop) of the two wings of FMN (Supp. Fig. 6C,D). Together, these results demonstrate that SMS-seq can capture secondary and tertiary dependencies, and the conformational changes of riboswitches upon ligand binding.

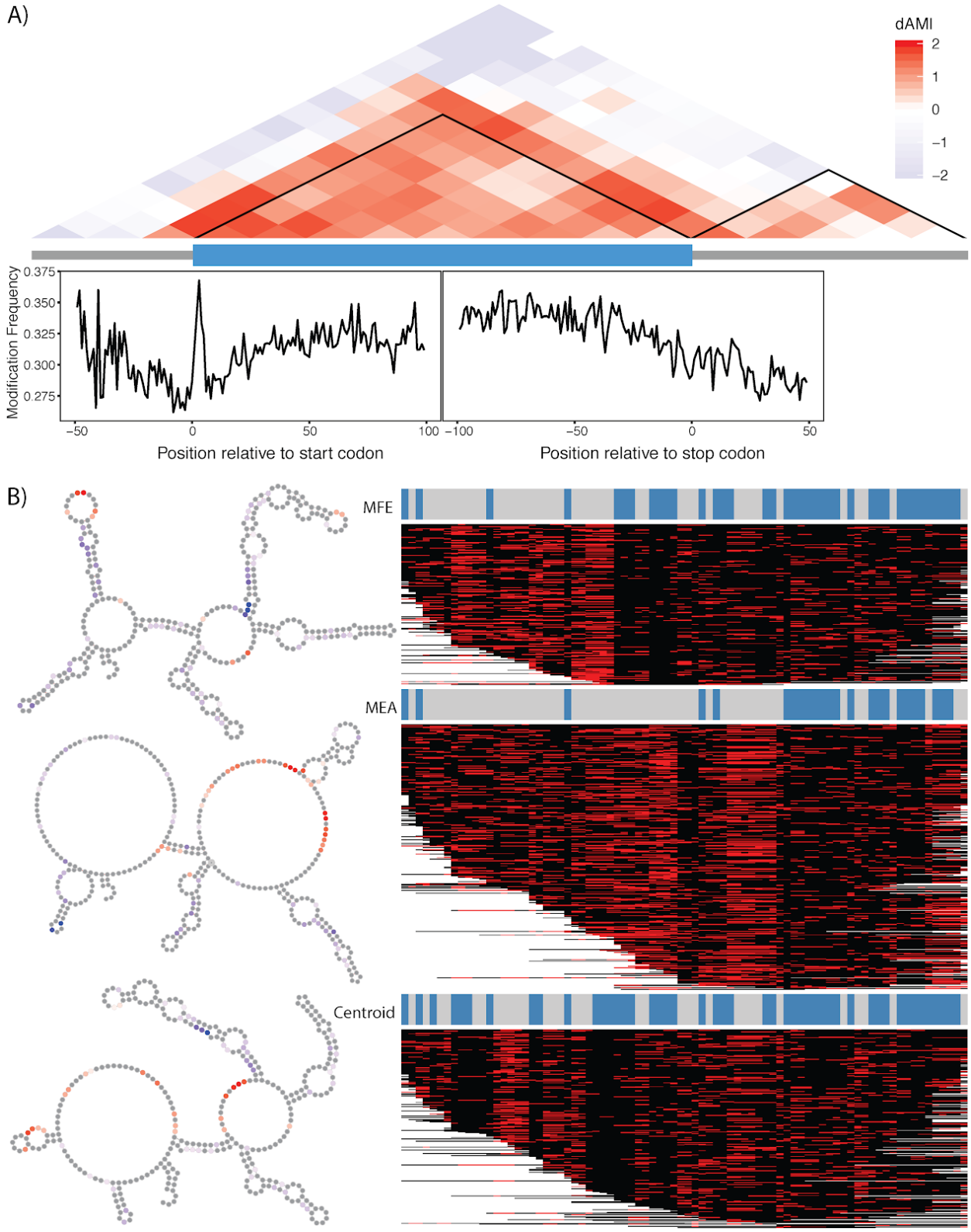

Fig 3. Structural probing of yeast messenger and non-coding RNA

A) The upper triangle shows average dAMI dependencies over all transcripts binned by position of the probed bases in their respective genes. The first two bins have been removed due to insufficient coverage. The lower panels show average modification rates at the boundaries of the coding region across all transcripts. **B)** Reads mapping to snR17a clustered by their distance to three *in silico* predicted structures.

Encouraged by the results on synthetic structures, we generated poly(A)-selected and non-poly(A) RNA libraries from yeast (*Saccharomyces cerevisiae*) and folded these *in vitro* in the presence and absence of DEPC. Average modification rates were highly reproducible across replicates (Pearson cc: 0.78, Supp. Fig. 8, Methods). When we compared to enzyme based PARS-assisted folding [13], we observed a significant difference in modification rates between nucleotides classified by PARS as open and closed (Supp. Fig. 9). Globally, our approach was able to confirm properties of mRNA structure showing an increase in modifications around the translation start site consistent with this region being open [13]; in contrast, the translation termination site and 3'UTR were less accessible (Fig. 3A, lower panels) as observed *in vivo* [14]. Over non-coding RNAs we observed agreement when clustering reads to known or predicted structures (Fig. 3B, Supp. Fig. 10, Methods). We could also observe known structural elements in mRNAs (Supp. Fig. 11) in line with previous study [14]. Together these findings confirmed that SMS-seq captures known patterns of mRNA structure.

We next asked whether mRNAs show signs of having general structural dependencies. We selected all mRNAs with high coverage (150 reads spanning CDS, Methods) and divided their CDS and UTRs into 10 and 5 bins, respectively. Due to limitations with the standard protocol for Nanopore direct RNA sequencing, 5' UTRs are frequently truncated and we therefore excluded

the start of transcripts from our analyses where coverage was low. For the remaining bins, we calculated the average dAMI between all pairs of bins (Fig. 3A, triangle, Methods). This analysis revealed that on average mRNAs have dependencies that are compartmentalized to the CDS and 3' UTR. Specifically, the CDS and 3' UTRs had a greater than expected dAMI relative to themselves than to each other. Structures in these regions are frequently associated with translation for the CDS and regulation for the 3' UTR [13,14]. Consistent with this, the CDS region with higher dAMI extends into both UTRs coinciding with regions involved in translation initiation and termination. This argues, at least *in vitro,* structure follows functional domains, and both CDSs and 3' UTRs have on average an internally more coordinated structure.

In summary, we introduced a method for single-molecule probing of RNA structure. SMS-seq does not require amplification or size selection, and does not need to be normalized based on position in the molecule – common sources of biases in previous methods [2]. Unlike the existing methods, reverse transcription in SMS-seq is optional thus avoiding another source of bias. Each molecule captures multiple modified bases and can therefore reveal heterogeneity of structures and dependencies within a structure. While dependencies require substantial coverage to obtain, we expect that with the rapid improvement to modification-calling algorithms the sensitivity of our approach will increase greatly and eventually be used as constraints for RNA structure prediction algorithms. Further improvements are also expected through different modifying reagents which could potentially be used *in vivo* that target all 4 nucleotides and have a strong effect on current signal in the nanopore. While we currently exclude most of the 5' UTR analysis, the method can reach full-length coverage of mRNAs through advances in the direct

RNA sequencing protocol and mapping algorithms. We expect full-length structural probing of native RNA by direct RNA sequencing will soon become a standard tool in predicting secondary structure and inferring structural dependencies and tertiary interactions.

MATERIALS AND METHODS

Sample preparation

In vitro synthesis of RNA. Custom sense and antisense hairpin oligos with T7 promoter sequence were purchased from Sigma (Sigma-Aldrich) and *in vitro* transcribed using T7 RNA polymerase (NEB). *F. nucleatum* FMN riboswitch RNA was synthesised *in vitro* using T7 RNA polymerase from a pUT7 vector that carries 112-nucleotides FMN sensing domain with hammerhead ribozyme at the end to obtain identical 3' end [11]. Coenzyme thiamine pyrophosphate binding riboswitch (TPP) RNA was synthesized *in vitro* using T7 RNA polymerase. TPP template was prepared using sense and antisense TPP oligos with T7 promoter sequence (Supp. Table 1) purchased from Sigma (Sigma-Aldrich) that was assembled using PCR, cloned to pCR 2.1-TOPO vector (Invitrogen) and digested using EcoR I (NEB).

Cell culture. Yeast (*Saccharomyces cerevisiae*) strain Y187 was grown in YPD medium at 30 °C. Saturated cultures were diluted to OD600 of ~0.1 and grown to a final OD600 of 0.7 followed by RNA extraction using hot phenol-chloroform method. Poly(A) enrichment was performed using Poly(A)Purist MAG Kit (Invitrogen) following manufacturer protocol. The remaining non-poly(A) RNA was precipitated overnight at -80 °C following RNA ethanol precipitation protocol.

Diethyl Pyrocarbonate (DEPC) treatment of RNA. All structural probing was carried out using DEPC (Sigma), which carbethoxylates unpaired adenosine at N-6 or N-7 by opening the imidazole ring [6]. 1µg RNA was denatured at 90 °C for 3 minutes, cooled on ice and renatured in 200 mM Hepes pH 7.8, 100 mM KCl, 10 mM MgCl2 for 20 minutes at 20 °C. For (+)DEPC, RNA was treated with 10% (final volume) DEPC for 45 min at 20 °C. DEPC reaction was performed in the absence and presence of sixfold molar excess of riboflavin 5'-monophosphate and thiamine pyrophosphate (Sigma-Aldrich) for FMN and TPP riboswitch, respectively. The reactions were immediately purified using RNA Clean & Concentrator (Zymo Research) or RNA XP bead (Beckman Coulter).

Oxford nanopore sequencing and mapping

Library preparation. Sequencing libraries were prepared following direct RNA sequencing protocol (Oxford Nanopore Technologies, ONT) by replacing the 3' adapter (RTA) with custom made barcodes for multiplexing (Supp. Table 1)[15]. After renaturation and treatment with DEPC, *in vitro* synthesized and yeast non-poly(A) RNA were polyadenylated using E.coli polyA polymerase (NEB) at 37 °C for 15 min. To obtain full length reads and a molecule long enough to be processed, the hairpin and riboswitches were ligated to the 3' end of an RNA (Supp. Table 1). RTA adapter ligated RNAs were reverse-transcribed using superscript III. Different barcodes were used for the (+)DEPC and (−)DEPC libraries. The first-strand complementary DNAs were pooled and a motor protein ligated, which were subjected to Oxford Nanopore sequencing using MinION Flow Cell (R9.4).

Read mapping. Reads were demultiplexed using poreplex (v 0.5)[15] and resquiggled using tombo v1.5 (ONT)[7] by mapped to their respective references. Yeast reference transcriptome was obtained from Ensembl (version R64-1-1). Reads were mapped by allowing mismatches, insertions and deletions to account for nanopore sequencing errors using minimap2 (v 2.9)[16].

Detection of DEPC modification

Chemical probing of adenine residues with DEPC. Several studies indicate that DEPC can be used for detecting single-stranded adenosine in RNA molecules [3–5,17–19]. Chemical reactivity is generally indicative of an absence of base stacking. Lack of reactivity, however, is more difficult to interpret since it could result not only from base pairing and tertiary interactions but also from the presence of stabilizing divalent cations [17].

We compared the (+)DEPC and (−)DEPC data sets and derived the final DEPC modification for adenosine using Tombo[7] method model_sample_compare with --fishers-method-context 3. Modification status of a nucleotide was assigned by comparing the current signal distribution between (+)DEPC and (−)DEPC samples. The tombo statistic at a level of 0.2 was used as proxy to determine the modification state of each adenosine. For determining the accuracy of modification calling on the hairpin (ROC curves) we excluded reads with <4 and >18 modifications, which we deemed to be under- and over-modified, respectively (Supp Fig. 3).

Inferring RNA structures in mRNA

Global transcriptome features. We annotated 5'UTRs of transcripts from *Saccharomyces*

cerevisiae reference genome (R64-1-1.79) using CAGE data (GSE69384) from Weyr et al.[20]. The transcription start site was defined as the largest CAGE peak (after normalization with CAGEr) [21] upstream of the transcript. We selected transcripts with good coverage (>150 reads, Supp. Fig. 8) that spanned the whole CDS to determine global mRNA features in yeast. For this and subsequent dependency analyses, 10 nts from the 5′ were excluded to avoid alignment problems because of unreliable current signals. In our analysis average DEPC modification of a region is the mean of the DEPC modification of all adenine nucleotides in that region, for all of the transcripts under consideration.

PARS scores were downloaded from http://genie.weizmann.ac.il/pubs/PARS10[13] for yeast transcriptome and compared with DEPC modification level. Modification level of known RNA structure elements of some transcripts were compared with PARS score.

Determination of dependencies

Distance adjusted mutual information (dAMI). Adjusted mutual information is a variation of mutual information correcting for agreement due to chance (scikit-learn 0.22.1)[22] which can be calculated for pairs of nucleotides to see how informative knowing the modification state of one base is on predicting the state of the other base. Due to the high dependency of bases in close proximity, we normalize the AMI score between two bases by the expected (mean) score calculated from all bases spaced at the same distance in yeast mRNA (Supp. Fig. 5). The log2 ratio of the AMI and the expected AMI we call dAMI. For the mRNA bins, dAMI scores were

calculated by taking the mean dAMI across all pairs of bases mapping to those two bins across all reads.

Deciphering RNA structural heterogeneity

The hairpin RNA and snoRNAs were folded using the ViennaRNA Package [9]. To determine hairpin heterogeneity, all reads were folded *in silico* with modifications used as hard constraints (not allowed to basepair). The number of reads resulting in the same structure were counted. Yeast non-coding RNAs were folded – based on sequence minimum free energy (MFE), maximum expected accuracy (MEA) and centroid – and reads were clustered to them using the binary distance metric i.e., the proportion of reads in which *both* of the two positions are modified/open relative to the number of reads where *at least one* is modified/open. The structures of the riboswitches were obtained from previous studies [10,11].

Data availability

The sequencing files are available on the European Nucleotide Archive website, under accession code PRJEB36658.

Code availability

Scripts used to process the data are available upon request.

REFERENCES

1. Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics* vol. 15 469–479 (2014).
2. Strobel, E. J., Yu, A. M. & Lucks, J. B. High-throughput determination of RNA structures. *Nature Reviews Genetics* vol. 19 615–634 (2018).
3. Rairkar, A., Rubino, H. M. & Lockard, R. E. Chemical probing of adenine residues within the secondary structure of rabbit 18S ribosomal RNA. *Biochemistry* **27**, 582–592 (1988).
4. Brunel, C., Romby, P., Westhof, E., Ehresmann, C. & Ehresmann, B. Three-dimensional model of Escherichia coli ribosomal 5 S RNA as deduced from structure probing in solution and computer modeling. *J. Mol. Biol.* **221**, 293–308 (1991).
5. Reineke, L. C., Komar, A. A., Caprara, M. G. & Merrick, W. C. A small stem loop element directs internal initiation of the URE2 internal ribosome entry site in Saccharomyces cerevisiae. *J. Biol. Chem.* **283**, 19011–19025 (2008).
6. Leonard, N. J., McDonald, J. J., Henderson, R. E. & Reichmann, M. E. Reaction of diethyl pyrocarbonate with nucleic acid components. Adenosine. *Biochemistry* **10**, 3335–3342 (1971).
7. Oxford Nanopore Technologies. Tombo: detection of non-standard nucleotides using the genome-resolved raw nanopore signal. https://nanoporetech.com/resource-centre/posters/tombo-detection-non-standard-nucleotides-using-genome-resolved-raw-nanopore (2017).
8. Homan, P. J. *et al.* Single-molecule correlated chemical probing of RNA. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13858–13863 (2014).

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

10. Serganov, A., Polonskaia, A., Phan, A. T., Breaker, R. R. & Patel, D. J. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441**, 1167–1171 (2006).

11. Serganov, A., Huang, L. & Patel, D. J. Coenzyme recognition and gene regulation by a flavin mononucleotide riboswitch. *Nature* **458**, 233–237 (2009).

12. Winkler, W., Nahvi, A. & Breaker, R. R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952–956 (2002).

13. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103–107 (2010).

14. 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).

15. Chang, H. Poreplex. *github* https://github.com/hyeshik/poreplex (2018).

16. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

17. Peattie, D. A. & Gilbert, W. Chemical probes for higher-order structure in RNA. *Proceedings of the National Academy of Sciences* vol. 77 4679–4682 (1980).

18. Moazed, D., Van Stolk, B. J., Douthwaite, S. & Noller, H. F. Interconversion of active and inactive 30 S ribosomal subunits is accompanied by a conformational change in the decoding region of 16 S rRNA. *J. Mol. Biol.* **191**, 483–493 (1986).

19. Van Stolk, B. J. & Noller, H. F. Chemical probing of conformation in large RNA molecules.

Journal of Molecular Biology vol. 180 151–177 (1984).

20. Wery, M. *et al.* Nonsense-Mediated Decay Restricts LncRNA Levels in Yeast Unless Blocked by Double-Stranded RNA Structure. *Mol. Cell* **61**, 379–392 (2016).

21. Haberle, V., Forrest, A. R. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* **43**, e51 (2015).

22. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).