Transcriptome-Wide Combinatorial RNA Structure Probing in Living Cells

Dalen Chan[1,$], Chao Feng[1,$], Whitney England[1,$], Dana Wyman[2], Ryan A. Flynn[3], Xiuye Wang[4], Yongsheng Shi[4], Ali Mortazavi[2], and Robert C. Spitale[*,1,5.]

(1) Department of Pharmaceutical Sciences, University of California, Irvine. Irvine, California. 92697, (2) Department of Developmental and Cellular Biology, University of California, Irvine. Irvine, California. 92697 (3) Department of Chemistry, Stanford University, Stanford CA 94305 (4) Department Microbiology and Molecular Genetics, University of California, Irvine. Irvine, California. 92697 (5) Department of Chemistry, University of California, Irvine. Irvine, California. 92697

($) These authors contributed equally to this manuscript.

*Correspondence: rspitale@uci.edu

Abstract

RNA molecules can fold into complex structures and interact with trans-acting factors to control their biology. Recent methods have been focused on developing novel tools to measure RNA structure transcriptome-wide, but their utility to study and predict RNA-protein interactions or RNA processing has been limited thus far. Here, we extend these studies with the first transcriptome-wide mapping method for cataloging RNA solvent accessibility, icLASER. By combining solvent accessibility (icLASER) with RNA flexibility (icSHAPE) data, we efficiently predict RNA-protein interactions transcriptome-wide and catalog RNA polyadenylation sites by RNA structure alone. These studies showcase the power of designing novel chemical approaches to studying RNA biology. Further, our study exemplifies merging complementary methods to measure RNA structure inside cells and its utility for predicting transcriptome-wide interactions that are critical for control of and regulation by RNA structure. We envision such approaches can be applied to studying different cell types or cells under varying conditions, using RNA structure and footprinting to characterize cellular interactions and processing involving RNA.

38 **Introduction**

39 Precise structural conformations are a hallmark of functional nucleic acid polymers in cells.
40 For example, the genome is arranged in a compact three-dimensional structure, whose dynamics
41 and solvent accessibility are critical for the control of gene expression(1). Methods to measure
42 these structural properties of the genome are now quite mature and can be done genome-wide
43 to understand the interactions between DNA and proteins, and even predict RNA expression
44 outcomes(2). As a corollary, there has also been a more recent emergence of approaches to
45 measure transcriptome-wide RNA structure in cells(3), but these efforts still lag behind in their
46 utility and impact, when compared to the biological significance of DNA-centric measurements
47 discussed above.

48 The ability of RNA molecules to fold into complex two- and three-dimensional structures
49 is critical for the many biological functions they perform(4, 5). Currently, the majority of reagents
50 used for structure probing can identify base-paired residues within RNA. For example, adenosine
51 and cytosinee residues not involved in Watson-Crick-Franklin pairing can be alkylated by
52 dimethylsulfate(6). Similarly, glyoxals have recently been observed to react with guanosine
53 residues at N-1 and exocyclic N-2 position(7). Selective hydroxyl acylation analyzed by primer
54 extension, or SHAPE, approximates 2'-hydroxyl flexibilities as a proxy for single-strandedness of
55 a residue(8, 9). These long-standing efforts demonstrate the power of regent design to understand
56 and relate chemical reactivity to RNA structure.

57 Transferring conventional chemical probes from one-RNA-at-a-time measurements to
58 transcriptome-wide scale has proven to be a formidable challenge. Bifunctional RNA structure
59 probing reagents that enable reactivity with RNA and enrichment of reaction sites has been
60 demonstrated to dramatically increase signal-to-noise ratios of RT-stops arising from chemical
61 adducts(10, 11). We previously developed icSHAPE, which features a bifunctional chemical probe
62 to measure nucleobase flexibility through RNA hydroxyl acylation transcriptome wide(10) (**Fig. 1,**
63 **a**). Solvent accessibility, an orthogonal property of RNA, has been more difficult to probe inside

cells, requiring the use of a synchrotron radiation light source(12). We recently described the development of LASER, or Light Activated Structural Examination of RNA(13). LASER relies on the facile light activation of a nicotinoyl azide to a nitrenium ion, which undergoes electrophilic aromatic substitution with electron rich purine residues at their C-8 nucleobase position. While this initial work demonstrated the utility of LASER to read out solvent accessibility, it still remained a critical challenge to extend these efforts to a transcriptome-wide scale and evaluate this approach to understand RNA structure and function more broadly.

Herein we describe our efforts to greatly expand the toolbox of RNA structure probing reagents, and demonstrate their utility to understand RNA structure in cells as well as interactions involving RNA to control RNA processing and RNA-protein interactions. We designed, synthesized, and implemented a bifunctional LASER probe to measure transcriptome-wide purine C-8 solvent accessibility (**Fig. 1, b,** *in vivo* click LASER, icLASER). Our design employed an alky azide functional group for enrichment, therefore permitting a common biochemical and computational approach to identify RNA-reagent adducts for measuring RNA structure (**Fig. 1, c**). By implementing this novel bi-functional probe, icLASER provides the first transcriptome-wide measurement of solvent accessibility. Towards the goal of integrating multiple, orthogonal measurements of RNA structure, we develop an approach to directly compare icSHAPE (hydroxyl acylation; flexibility) and icLASER (solvent accessibility). We highlight the power these data and this strategy in the context of predicting RNA-protein interactions and RNA polyadenylation. Our results demonstrate that combinatorial RNA structure probing can be employed to better understand RNA structure and processing in cells transcriptome-wide, and is thus a highly-useful complement and/or predictor of critical interactions that control RNA biology.

Results

Development of a bifunctional LASER probe

An optimal probe for icLASER would retain the selectivity of NAz and also allow for enrichment using conditions non-harmful to RNA. As such, NAz-N_3, which we predicted would be able to undergo light transformation to the activated nitrenium ion, would also preserve the alkyl azide for Cu-free "click" after RNA adduct formation. Notably, aroyl azides are amenable to activation by lower energy long wavelength light, whereas alkyl azides need higher energy short wavelength light for activation(14-16). We hypothesized these special characteristics of azide stability would permit photo-specific activation of NAz-N_3 for icLASER probing (**Fig. 2, a**).

To test our hypotheses, we synthesized NAz-N_3 and compounds 1 and 2 (**Fig. 2, b; Supplementary Information**). The NAz-N_3 probe is synthesized efficiently from commercially available methyl 6-methylnicotinate in six steps in 43% overall yield. Treatment of methyl 6-methylnicotinate with trichloroisocyanuric acid generated monochloro substitution on the methyl group selectively, followed by alkaline hydrolysis of the ester to afford the corresponding carboxylic acid. The acetyl azide is installed routinely by transformation of the carboxylic acid into acetyl chloride using oxalyl chloride and azide substitution. The above monochloro substituent is finally converted to iodide through a Finkelstein reaction, which is replaced subsequently by azide to furnish NAz-N_3. Synthesis of 1 and 2 is reported in **Supplementary Information**.

We obtained UV-Vis spectra of these compounds, to understand differences in their absorbance properties (**Fig.2, c**), and observed that only NAz-N_3 and **2** have absorption in the long wavelength UV region used previously by us in LASER and known to be specific for aroyl azides (**Fig.2, d&e**)(13). Consistent with what others have observed with azide-specific activation, these data suggest that light activation of NAz-N_3 should be specific to the aroyl azide. To test reactivity and preservation of a SPAAC-compatible azide, we incubated an *in vitro* transcribed SAM-I riboswitch with NAz-N_3 and exposed the solution to (+/-) long wavelength UV light. Following irradiation, we isolated the RNA and incubated the RNA with dibenzylcyclooctyne-

116 conjugated to biotin (DBCO-biotin). We observed biotin conjugation on a streptavidin dot blot only
117 in the sample incubated with long wavelength UV light and NAz-N_3 (**Fig.2, f**).
118 We next tested if NAz-N_3 enriched RNA would produce similar truncation profiles to NAz
119 probing (LASER). We used NAz-N_3 modified riboswitch RNA, biotinylated the RNA using SPAAC,
120 and subjected it to streptavidin-coated magnetic bead enrichment. Biotinylated RNA was then
121 eluted and compared against input RNA. Each sample was then reverse transcribed with a ^{32}P-
122 cDNA primer and analyzed by denaturing gel electrophoresis. As shown in **Fig. 2, g,** the NAz-N_3
123 enriched samples displayed significantly higher signal for structure stop cDNA and dramatically
124 reduced full-length extensions. Finally, NAz-N_3 elution mapped onto sites of modification similar
125 observed with NAz (A & G residues; **Fig. 2, h**). Further, we also demonstrated that NAz-N_3 can
126 modify RNA and measure RNA structure in cells, and observed differences *in vitro* and in cells
127 that correspond to cDNA profiles to that of the parent compound NAz(17) (**Supplementary Figure**
128 **1**). Overall, this result demonstrates that NAz-N_3 is capable of enrichment using SPAAC reactions,
129 yet still yields consistent truncation signatures for structural analysis.
130
131 **Transcriptome-wide implementation of NAz-N_3**
132 We modified K562 cells (*in vivo*) and K562 total RNA isolated and re-folded outside of cells
133 (*in vitro*) in independent experiments with NAz-N_3 (icLASER) and NAI-N_3 (icSHAPE). We applied
134 our well-established protocol for the isolation of poly-adenylated (polyA) RNA and mapping of
135 biotin-conjugated RNA structure probes through deep sequencing (**Figure 1, Supplementary**
136 **Methods)** (18, 19), which generated more than 100 million reads for each dataset
137 (**Supplementary Figure 2**). Each dataset had strong agreement with each other and RT stops
138 were enriched for G and A.
139 Mapping RT stops from icLASER enriched libraries revealed a marked enrichment of A
140 and G residues, which was expected given the reactivity profile of the NAz-N_3 probe
141 (**Supplementary Figures 2 & 3**). This profile was unique to icLASER, whereas for icSHAPE we

142 observed a slight enrichment for adenosine stops, as seen previously (**Supplementary Figure**
143 **3**) (10). Importantly, we find a strong correlation between +UV samples, whereas, the distribution
144 of +UV and –UV RT-stops showed very low correlation (**Supplementary Figure 4**). Comparison
145 of the icSHAPE and icLASER stops showed strong intra-reagent correlation but weaker inter-
146 reagent correlation, suggesting each probe measured a discrete set of positions in the
147 transcriptome.

148 As an initial quality control, we analyzed the icLASER and icSHAPE data against manual
149 footprinting on an RNA structure in which we had reasonably high-resolution data: the ribosome.
150 We focused on a section of 18S rRNA has been previously interrogated by footprinting(13) (**Fig.**
151 **3, a & b**). The manual footprinting for NAI-N_3 in this region showed some slight differences
152 between in and outside of cells (**Fig. 3, c**). In similar fashion, our icSHAPE data had similar overall
153 reactivity between the two experimental setups (**Fig. 3, d**). Analysis of the RNA in this region (Fig.
154 3, e) revealed that it was solvent protected by rRNA and ribosome binding proteins, but was
155 largely still single stranded and thus would be reactive toward SHAPE reagents. We also
156 performed comparison probing with NAz-N_3 on the same stretch of rRNA and observed marked
157 differences between in and outside the cells (**Fig. 3, f**). Consistent with this analysis was the
158 reactivity profile for icLASER (**Fig. 3, g**), which is supported by the solvent protected nature of
159 this stretch of RNA in the 18S ribosome CryoEM model (**Fig. 3, h**). Overall, these results suggest
160 that icLASER and icSHAPE are both measuring environment-unique aspects of RNA structure in
161 and outside cells.

162 As this is the first transcriptome-wide map of solvent accessibility probing with icLASER,
163 we also analyzed mRNA profiles. We observed that start codons are largely open and solvent
164 accessible with high reactivity at the A and G of the AUG (**Supplementary Figure 5**). The stop
165 codon also displayed high reactivity in the last two positions (**Supplementary Figure 5**), and both
166 of these observations are in line with previous measurements using icSHAPE and DMS-seq
167 (20).To understand what regions of icLASER and icSHAPE signal could be due to conditions

168 specific to living cells, we compared the *in vivo* and *in vitro* profiles as previously described for
169 icSHAPE (VTD = *in vivo* profile – *in vitro* profile, (21)). Specifically, we examined all hexamers
170 across the transcriptome, and consistent with our previous reports, icSHAPE structure was overall
171 higher *in vivo* (**Supplementary Figure 6.**). This is hypothesized to be correlated with structural
172 remodeling *in vivo*, in comparison to *in vitro* (10). In contrast, we observed a bimodal distribution
173 of solvent accessibility differences in and outside of cells with icLASER. Hexamers that include
174 the Kozak sequence showed small VTD differences, which is in line with previous reports with
175 icSHAPE(21). However, for icLASER, hexamers containing RNA binding protein motifs had lower
176 maximal averaged VTD values. This observation may be consistent with the known reactivity of
177 LASER probing, which can result in a change in chemical reactivity in cells due to protein
178 footprinting, suggesting that icLASER could be used to map protein-RNA interactions. Overall,
179 these results demonstrate the robustness of icLASER probing and demonstrate its utility in
180 mapping the structures of RNAs in and outside of cells.
181
182 **Global prediction of protein footprints by combining structure probing tools.**
183 Despite the importance of RNA-protein interactions and the interplay of RNA structure and
184 protein binding, there has thus far been a lack of transcriptome-wide analysis of RNA structure
185 motifs with the express goal of attempting to predict binding sites. To address this gap, we
186 assessed if icLASER, icSHAPE, or their combination could be used to predict RNA-protein
187 interactions (**Fig. 4**). As a reference for experimentally measured RNA binding protein (RBP) sites,
188 we leveraged enhanced individual nucleotide resolution CLIP (eCLIP) datasets from ENCODE
189 collected from K562 cells (22, 23). We focused on RBPs that had significant binding in 3'-UTRs
190 (ensuring high sequence coverage in icLASER and icSHAPE data), binding motifs with at least
191 three A or G residues, targets expressed at greater than or equal to 5 RPKM. As control regions,
192 we sampled the same number of sites with matching RNA motifs not called as peaks in the eCLIP
193 data.

194 We first focused on PUM2, a well-established RNA binding protein that is important in
195 regulating RNA localization and translation, and has high-resolution structures for comparison
196 (**Fig. 4, a**)(24). Comparative icLASER and icSHAPE profiles in and outside of cells yielded key
197 differences that likely reflect protein binding (**Fig. 4, b**). First, the G1 position has much higher
198 icLASER signal *in vivo* and this position has its C-8 atom pointed directly out into solvent,
199 suggesting that the protein may be holding this portion of the RNA in this conformation. Positions
200 A3-5 have minimal differences *in vivo* and *in vitro* by icLASER and these positions are also out
201 away from the structure. In contrast, the icSHAPE profile has lower reactivity at position A3, and
202 this may be due to different levels of stability from binding. These differences overall, and the
203 reactivity differences observed at RNA binding sites overall (**Supplementary Figure 6**),
204 suggested to us that structure probing using these reagents could be utilized to predict RNA-
205 protein interactions. To test this, we employed support vector machine learning (SVM)(10),
206 training the classifier using the VTD profiles at eCLIP-defined binding sites *in vitro* and *in vivo*.
207 We trained the classifier using icSHAPE, icLASER, or combined profiles, using randomly selected
208 sites as a training set. We evaluated performance using the area under the reporter operator
209 characteristic (ROC) curve for PUM2 and observed that combining icSHAPE and icLASER
210 profiled dramatically increased the ROC area and predictability. (**Fig. 4, c)**. These results suggest
211 that a larger analysis of comparative structure probing could enable more analyses and prediction
212 of RNA-protein interactions in cells.

213 We expanded our initial analysis to other RBPs that have reported crystal or NMR
214 structures bound to cognate RNA sequences. RBFOX2 (**Fig. 4, d&e**) displayed a similar icSHAPE
215 structure profile to our previous efforts with icSHAPE in mouse embryonic stem cells(21).
216 icLASER profile differences showed high reactivity at the G1 position, and correspondingly the C-
217 8 of G1 is pointed out to solvent in the crystal structure. The same is true for A3, where the C-8
218 position is solvent exposed. Quaking1 (QK1) (**Fig. 4, f&g**) shows generally lower icSHAPE
219 reactivity across its binding motif. These positions are also protected from solvent either through

220 direct protein interactions or shielding from phosphate backbone, as is the case for A3 in the motif.
221 Lastly, TRA2A (**Fig. 4, h&i**) displayed correlation at positions 1 and 2 (AA); these are both solvent
222 exposed and likely to be highly dynamic, and as such had high icLASER and icSHAPE reactivity.
223 The last two A residues were protected from solvent by protein interactions and would likely be
224 stabilized to limit their SHAPE reactivity. Overall, these data show icLASER and icSHAPE provide
225 non-overlapping nucleotide resolution at RBP binding sites that can read out the biophysical
226 conformation of the bound RNA in a manner that corresponds to each method's chemical
227 properties.

228 To further expand our analysis of combined chemical probing we evaluated performance
229 using the area under the reporter operator characteristic (ROC) curve for each protein (**Fig. 4j**).
230 Using this approach, icLASER and icSHAPE data alone were able to predict between 50-70% of
231 the binding sites for 75 RBPs analyzed herein. Overall, the predictability was higher for icLASER
232 in comparison to icSHAPE. However, when we combined the icLASER and icSHAPE datasets,
233 we observed a dramatic increase in predictive power for protein binding sites and were now able
234 to predict over 85% protein occupancy on RNAs. We suspect the robust, but in some instances
235 imperfect, prediction may be due to the inability of LASER probing to differentiate RNA protection
236 from proteins versus RNA tertiary structure. It is worth noting that for 75 RBPs (300 experiments:
237 biological duplicate eCLIP + biological duplicate background) we are predicting their binding with
238 8 experiments (icSHAPE/icLASER biological duplicate + biological duplicate background).
239 Overall, these results support the notions that: (1) RNA solvent accessibility probing can be
240 utilized to predict RNA-protein interactions, (2) icLASER data can be utilized in conjunction with
241 CLIP datasets to further support protein occupancies determined by orthogonal methods, and (3)
242 by combining icLASER and icSHAPE, robust and transcriptome-wide predictions of many RBPs
243 is possible without protein-centric techniques such as CLIP.
244
245

246 **Predicting polyA sites in RNA with structure probing**

247 Analysis discussed previously (in relation to VTD; **Supplementary Figure 6**) revealed
248 hexameric sequences that have differences inside and outside of cells. One such sequence that
249 had higher in-cell reactivity for icLASER (and icSHAPE) was the polyadenylation signal sequence
250 (PAS) within the 3'-UTRs(25). Given the measurable VTD value, we explored if our structure
251 probing data could be used predict polyadenylation transcriptome-wide. This would be an
252 extremely important further validation of our approach to use structure probing for predicting RNA-
253 protein interactions and potential processing events.

254 To experimentally determine PAS sites, for SVM analysis, we generated the first
255 polyadenylation sequencing data (PAS-seq (26)) for K562 cells. PAS-seq uses polyA tail priming
256 to identify the sites of polyA tail selection (**Fig. 5, a**)**.** Inspection of PAS reads demonstrated clear
257 read buildup at the 3'-end of transcripts (**Fig. 5, b**) and we obtained sites of high and low PAS
258 read depth (**Fig. 5, c**). We then compared the icLASER and icSHAPE profiles at the PAS sites
259 and sequences with the same motif which are not annotated to be PAS sequences (negative
260 control) (27). We noticed a striking structural difference between them: a large peak at the first
261 two adenosine residues, followed by a drop in icLASER signal for the remaining UAAA (**Fig. 5,**
262 **d**). To understand if these differences were related to a biophysical conformation of the RNA in
263 an active PAS-recognition complex, we examined a newly published structure(28). This structure
264 contains three large proteins which form into a complex and recognize the AAUAAA motif (**Fig. 5,**
265 **e**). Close inspection of the structure revealed that the two five prime adenosine residues are
266 completely solvent exposed, whereas the UAAA residues are solvent protected by protein binding.
267 Impressively, the two AA residues have their C-8 positions (site of icLASER reaction) completely
268 exposed (**Fig. 5, e**). Consistent with our icLASER data, sites with high icLASER signal at the first
269 two adenosine residues had very high coverage by PAS-seq. This data further suggested to us
270 that icLASER signal alone could be used to predict polyA site selection. To test this hypothesis,
271 we utilized SVM(10) and demonstrated that icLASER and icSHAPE used together had an AUC

272 of 0.9 for predicting PAS sites (**Fig. 5, f**). These data nicely demonstrate that icLASER (and
273 combined icLASER and icSHAPE) can be used to predict sites of posttranscriptional regulation
274 and could be integrated with orthogonal datasets to interpret posttranscriptional processing of
275 RNAs.
276
277 **Discussion**
278 Here, by developing novel RNA structure bi-functional probes, we extend the utility and
279 flexibility of our previously reported RNA solvent accessibility probe, NAz. The new reagent, NAz-
280 N_3, similarly relies on specific photoactivation of an aroyl azide by long-wavelength UV light, but
281 can subsequently be ligated to biotin using copper-free "click" chemistry, for enrichment of
282 modified sites of adduct formation. Using this chemistry, we measured RNA solvent accessibility
283 in K562 cells, transcriptome-wide.
284 With the development of transcriptome-wide RNA structure probing techniques, an
285 exciting but thus far poorly explored aspect of these data has been the possibility to infer
286 interactions between RNAs and their cellular partners. By employing a computational strategy
287 (SVM) to combine icSHAPE and icLASER (as well as other reactivity-based measurements) we
288 take a critical step towards learning the potential of these methods in predicting RBP binding and
289 other functional RNA processing activities like PAS selection. We demonstrate the power of this
290 approach by predicting RNA-protein binding sites for a large number of RNA binding proteins. We
291 demonstrate that such an approach could be very powerful for measuring RNA-protein interfaces
292 and is highly complementary with protein-centric methods such as eCLIP. Further, we utilize RNA
293 structure probing to identify a structure signature associated with polyadenylation sites, which is
294 due to the presence of a protein bound at the PAS site. This extension also enables structure
295 probing to be utilized for other aspects of RNA biology, such as RNA processing.
296 The complexity of nucleic acid polymer structure is now well appreciated for DNA and the
297 folding of the genome; importantly, measuring the structure of such nucleic acids and how proteins

interact with DNA has been incredibly valuable for understanding how the genome is regulated to control biological processes within cells. Our goal herein is to make headway toward this goal for RNA, by the continued development of tools for transcriptome-wide measurement of RNA structure and RNA interactions that can contribute to its biological function and regulation. We anticipate that icLASER (a new aspect of transcriptome-wide RNA structure probing via solvent accessibility) will become an increasingly useful chemical tool to probe RNA structure in living cells transcriptome-wide. By demonstrating the predictive power of these tools, we expect RNA structure probing to expand its value into aspects of RNA biology.

Contributions.

D.C. worked on generating biochemical data for structure probing and libraries for RNA sequencing. C.F. worked on chemical synthesis and biochemical characterization of all compounds. W.E. worked on data analysis and RNA structure analysis. R.F. helped with generating RNA sequencing libraries. D.W. and A.M. worked on data analysis and discussions on RNA-protein interactions. X.W. and Y.S. worked to help with generating PAS seq data and WE analyzed the PAS seq datasets.

Data availability.

icLASER and icSHAPE datasets are deposited on GEO under the accession number GSE132099.

Additionally, icSHAPE datasets are deposited on the ENCODE database under experiments ENCSR976RFC, ENCSR803XFA, ENCSR286LXS, ENCSR992XHC, ENCSR052BBY, and ENCSR836VQU

PAS sequencing datasets are deposited on GEO under the accession number GSE145400.

Acknowledgements.

We thank the Spitale lab for their careful reading of the manuscript. This research was supported by the NIH (5UM1HG009443-03 to RCS). RCS is a Pew Biomedical Scholar.

Online Methods.

All methods pertaining to the synthesis of all probes is presented in the Supplementary Information.

334 **Primers used for Reverse Transcription:**

335 SAM-I (5', ATTTAGGTGACACTATAGTT, 3')

336 18s Primer 1 (5', CCAATTACAGGGCCTCGAAA, 3')

337

338 **SAM-I Construct (Control RNA for analysis of NAz-N_3 structure probing):**

339 A 94 nucleotide construct consisting of the sequence for the SAM-I riboswitch from the metF-
340 metH2 operon of *T. tencongensis* was designed into a plasmid with IDT. The plasmid encoding
341 the SAM-I sequence was transformed into a One Shot Top 10 chemically competent cells
342 (ThermoFisher) and plated on lysogeny broth (LB, ThermoFischer) supplemented with 100
343 mg/mL ampicillin (VWR) agar plates. A single colony was selected in a 3 mL culture and grown
344 overnight. The resulting plasmid was isolated using a QIAprep Miniprep kit (Qiagen). Transcription
345 template was prepared by PCR in 50 μL volumes using primers directed against the T7 promoter
346 (5', TAATACGACTCACTATAGGG, 3') and an adaptor sequence for reverse transcription (5',
347 ATTTAGGTGACACTATAGTT, 3') with Phusion High-Fidelity PCR master mix (NEB). The
348 mixture was treated with 1 μL of DPN1 (NEB) restriction enzyme and incubated at 37 °C for 30
349 min. The resulting PCR product was purified using a QIAquick PCR purification kit (Qiagen). *In*
350 *vitro* transcription of SAM-I RNA was performed in 20 μL reactions using 800 ng of template SAM-
351 I DNA with a T7 Ribomax transcription kit (Promega). Samples were treated with 1 μL of RQ1
352 DNAse (Promega) at 37 °C for 30 min. The resulting solution was brought up to 100 μL and
353 precipitated with 10 μL of 3 M NaOAc (pH 5.2), 1 μL glycoblue (ThermoFisher), and 300 μL
354 ethanol. The resulting RNA stored at -80 °C for 30 minutes, centrifuged at 15,000 rpm for 15
355 minutes, and resuspended in 50 μL of RNase-free water. An aliquot of SAM-I was run on a
356 denaturing PAGE gel (15% polyacrylamide, 0.5X TBE, 7 M urea) alongside a 100 bp RNA ladder
357 (Invitrogen). The band of interest was visualized with 1x SYBR-gold stain (ThermoFisher) in water

358 for 15 minutes. Resulting concentrations of RNA was quantified by integrating intensity of the
359 ladder with the RNA band of interest.
360
361 **^{32}P End labeling for reverse transcription:**
362 Primer DNA was 5' end labeled in 10 μL reactions in a T4 PNK mix (1 μL 10x T4PNK reaction
363 buffer, 2 μL 100 mM DNA primer, 5 μL nuclease free water, 1 μL γ-^{32}P-ATP, 1 μL T4 PNK, NEB).
364 The reaction was allowed to proceed for two hours at 37 °C. Reactions were stopped by the
365 addition of 5 μL of Gel Loading Buffer II. The reaction was loaded onto a 15% denaturing PAGE
366 gel. The band of interest was visualized by a phosphorimager (Typhoon, GE healthcare). The
367 resulting band was excised and eluted overnight in 400 μL of 300 mM KCl. Resulting solution was
368 EtOH precipitated and dissolved to 8,000 counts per minute (cpm)/μL for further use in reverse
369 transcription.
370
371 **Modification of RNA *in vitro*:**
372 5 μg total RNA (isolated from HeLa cells) or 10 pmol of *in vitro* transcribed RNA in 6 μL metal-
373 free water was heated for 2 minutes at 95 °C. The RNA was then flash cooled on ice for 1 minute,
374 and brought to room temperature. 3 μL of 3x RNA folding buffer (333 mM HEPES, pH 8.0, 20 mM
375 $MgCl_2$ and 333 mM NaCl) was added, and the RNA was allowed to equilibrate at 37 °C for 5 min.
376 To this mixture 1 μL of 3 M NAz, 1 M NAz-N_3 in DMSO (+) or DMSO (-) was added. Reactions
377 were then exposed to 20 Watt lamp (Zilla Desert UVB 50) UV light for 3 minutes for NAz, and 10
378 minutes for NAz-N_3. Reactions were brought up to 200 μL water, and extracted once with 200 μL
379 acid phenol/chloroform/isoamyl alcohol (pH 4.5, Ambion), and washed twice with 200 μL
380 chloroform (Sigma). Samples were precipitated by adding 20 μL 3 M NaOAc (pH 5.2), 1 μL
381 glycoblue (20 μg/μL) and 600 μL EtOH. Pellets were washed twice with 70% cold ethanol and
382 resuspended in 5 μL nuclease free water.

383

384 **Copper free click Chemistry of Modified RNA:**

385 In a reaction volume of 50 μL, modified RNA (10 pmol) was incubated with 5 μL
386 Dibenzocyclooctyne-PEG4-biotin (DBCO-Biotin, 10 mM, Sigma) for 1 h at 37 °C in 1x PBS.
387 Reactions were brought up to 200 μL and extracted once with 200 μL acid
388 phenol/chloroform/isoamyl alcohol (pH 4.5), and washed twice with 200 μL chloroform. Samples
389 were precipitated by adding 20 μL of 3 M NaOAc (pH 5.2), 1 μL glycoblue (20 μg/μL) and 600 μL
390 EtOH. Pellets were washed twice with 70% cold ethanol and resuspended in 2 μL nuclease free
391 water for dot blot analysis, or 10 μL nuclease free water for streptavidin enrichment assays.

392

393 **Dot blot Analysis of Enriched modified RNA:**

394 Hybond N+ membranes (GE) were pre-incubated in 10X SSC. Precipitated biotinylated total RNA
395 was dissolved in 2 μL of RNase free water. RNA was loaded onto the Hybond membrane and
396 crosslinked using 254 nm ultraviolet light. The membrane was incubated with blocking solution
397 (120 mM NaCl, 16 mM Na_2HPO_4, 8 mM NaH_2PO_4, 170 mM SDS) for 30 min. To the membrane
398 was added 1 μL Pierce high sensitivity streptavidin-HRP(ThermoFisher) in blocking solution. The
399 membrane was washed twice with wash buffer A (1:10 blocking solution) for 30 min, and twice
400 with wash buffer B (100 mM Tris pH 9.5, 100 mM NaCl, 20 mM $MgCl_2$) for 5 min. Membrane was
401 incubated with Pierce Western blotting substrate and visualized on the ChemiDoc (Biorad) under
402 chemiluminesence hi sensitivity. To visualize RNA, the membrane was stained with a methylene
403 blue solution (0.2% w/v methylene blue, 0.4 M sodium acetate).

404

405 **Modification of RNA in cells:**

406 HeLa cells were grown in DMEM (4.5 g/L glucose & L-glutamine [-] sodium pyruvate,
407 Thermofisher) culture medium supplemented with 10% FBS (SAFC) and 1% penicillin

408 streptomycin (Life technologies). K562 cells were grown in RPMI 1640 supplemented with 10%
409 FBS, and 1% penicillin streptomycin. Cells were washed three times with Dulbecco's phosphate-
410 buffered saline (DPBS, Genesee) and centrifuged at 1000 RPM for 5 min. Cells (~3-6 x 10^7) were
411 resuspended in 45 μL DPBS. 5 μL DMSO (-), 10% final concentration, 5 μL 3 M NAz or 1 M NAz-
412 N_3 in DMSO (+) was added to the desired final concentration. Cell suspensions were subjected
413 to UV light for 5 minutes using NAz and 10 minutes with NAz-N_3. Cells were pelleted by
414 centrifugation at 1000 RPM for 5 min and resuspended in 1 mL Trizol Reagent (ThermoFisher).
415 RNA was harvested using Trizol Reagent following the manufacturer's instructions. 500 μL of
416 aqueous phase was then precipitated with 500 μL isopropyl alcohol at room temperature for 10
417 minutes. Samples were centrifugated at 15,000 RPM at 4 °C and washed twice with cold 70%
418 EtOH.
419
420 **Enrichment of Modified RNA:**
421 In 700 μL reaction volume, 50 pmol of biotinylated RNA was added with 50 μL of prewashed
422 Dynabeads MyOne C1 beads (ThermoFisher). The solution was then mixed at room temperature
423 for 1 h. The beads were collected on a magnetic plate and flowthrough was saved. The beads
424 were then washed three times with 700 μL of Biotin Wash buffer (10 mM Tris-HCl, pH 7.0, 1 mM
425 EDTA, 4 M NaCl, 0.2% Tween). The first wash was saved and combined with the flowthrough for
426 further analysis. Samples were later washed twice with RNase-free water. NAz-N_3 adducts
427 underwent harsher wash conditions was incubated twice with 700 μL Biotin wash buffer for 5 min
428 along with two washes with RNase-free water at 70 °C. Samples were eluted twice with 44 μL
429 formamide, 1 μL of 0.5 M EDTA, and 5 μL of 50 mM of free D-Biotin at 95 °C for 5 min. Eluted
430 samples were diluted with 600 μL RNase-free water. All samples were purified using RNA Clean
431 and Concentrator Kit (Zymo). Samples were eluted in 6 μL of RNase-free water and used for
432 subsequent reverse transcription.

433

434

435 **Reverse transcription of modified RNA (in vitro and in vivo):**

436 ^{32}P-end-labeled DNA primers were annealed to modified RNA by incubating 95 °C for two
437 minutes, then 25 °C for two minutes, and 4 °C for 2 minutes. To the reaction, 1 μL of 5x First
438 strand buffer, 0.5 μL nuclease free water, 0.5 μL 100 mM DTT, and 0.5 μL 10 mM dNTP's were
439 added. The reaction was preincubated at 52 °C for 1 min, then 1 μL superscript III (ThermoFisher)
440 was added. Extensions were performed for 15 min. To the reaction, 1 μL 4 M sodium hydroxide
441 was added and allowed to react for 5 min at 95 °C. The resulting complementary DNA (cDNA)
442 was snapped cooled on ice, and ethanol precipitated according to above procedures. Purified
443 cDNA was resuspended in 2 μL of nuclease-free water and 2 μL of Gel Loading Buffer II. cDNA
444 products were resolved on 10% denaturing polyacrylamide gel, and visualized by a gel imager
445 (Typhoon, GE healthcare).

446

447 **Generation of PolyA sequencing libraries.**

448 10 μg total RNA extracted with Trizol (Ambion) was fragmented with fragmentation reagent
449 (Ambion) at 70 °C for 10 minutes followed by precipitation with ethanol. Reverse transcription was
450 performed with PASSEQ7-2 RT oligo:

451

452 [phos]NNNNAGATCGGAAGAGCGTCGTGTTCGGATCCATTAGGATCCGAGACGTGTGCTCT
453 TCCGATCTTTTTTTTTTTTTTTTTTTTT[V-Q]

454

455 and Superscript III (Invitrogen). cDNA was recovered by ethanol precipitation and 120-200
456 nucleotides of cDNA was gel-purified from 8% Urea-PAGE. Recovered cDNA was circularized
457 with Circligase™ II (Epicentre) at 60 °C overnight. Buffer E (Promega) was added in cDNA and

458 heated at 95 °C for 2 minutes, and then cool to 37 °C slowly. Circularized cDNA was linearized
459 by BamH I (Promega). cDNA was collected by centrifugation after ethanol precipitation. PCR was
460 carried out with primers PE1.0 and PE2.0 containing index (Illumina). Around 200 bp of PCR
461 products was gel-purified and submitted for sequencing (single read 100 nucleotides).
462
463 **Data analysis for icSHAPE and icLASER**
464 Raw sequence reads were quality-checked using FastQC (29) and demultiplexed. Reads were
465 aligned to the ENSEMBL Release 88 GRCh38 transcriptome (30)and per-position enrichment
466 scores were calculated using the icSHAPE pipeline(21). Default parameters were used with the
467 exception of the final filtering step, where minimum values for hit coverage and background base
468 density were removed (-T 0 -t 0).
469
470 Publicly available eCLIP peak data for five RNA binding proteins in K562 cells was downloaded
471 from the ENCODE project(31). Known 5bp binding motifs BiorXiv
472 https://doi.org/10.1101/179648)) for each protein were located in each peak; peaks lacking the
473 motif were discarded. icLASER enrichment scores were extracted for each motif position, plus
474 five basepairs up- and downstream from the motif. Negative control sites were identified as
475 occurrences of the same 5bp motif that fall outside of eCLIP peaks. As with binding sites,
476 icLASER enrichment scores were extracted for a 15bp range centered on the motif. A number of
477 negative sites equal to the number of positive sites was selected randomly from the pool of
478 possible negatives.
479
480 icLASER analysis at poly-A signal sequences was conducted similarly; however, Ensembl-
481 annotated polyadenylation signal sequences were used to identify positive sites, and enrichment
482 scores were retrieved for a range of 20bp up- and downstream of the 6bp polyadenylation signal
483 motif.

484

485 **SVM analysis of icSHAPE/icLASER and RNA-binding proteins**

486 SVMs were implemented using libSVM (21). Both training and test data were scaled to a 0-1
487 scale and SVM parameters were selected using 5-fold cross-validation. For each motif, both *in*
488 *vitro* and *in vivo* data sets were divided in half at random; half for the data was used to train the
489 SVM, while the remaining half was used to test its predictions. icLASER scores at each position
490 in the 15bp range surrounding the motif of interest were input as features. Ranges containing null
491 values for icSHAPE or icLASER enrichment scores were discarded.

492

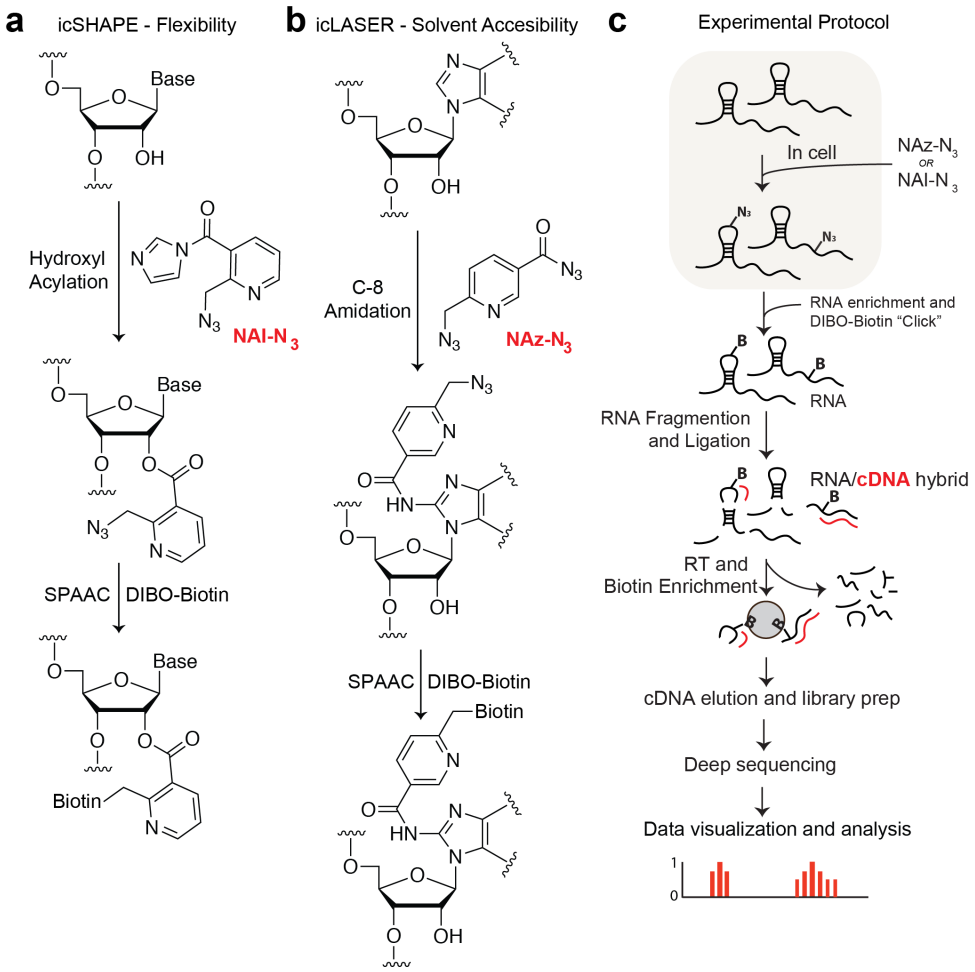
493 **SVM analysis of icSHAPE/icLASER and polyadenylation sites**

494 SVM analysis was conducted as with RNA-binding proteins. icSHAPE or icLASER scores for a
495 46bp range centered on the polyadenylation signal sequence were used as features.

496

497 **Figures.**

498



499

500

501 **Figure 1: Schematic for icSHAPE and icLASER. a.** Chemical protocol for icSHAPE(10). **b.**

502 Chemical protocol for icLASER. **c.** Protocol for library preparation for icSHAPE and icLASER,

503 which could be integrated in the same protocol method for RNA structure probing, transcriptome-

504 wide.

505

506

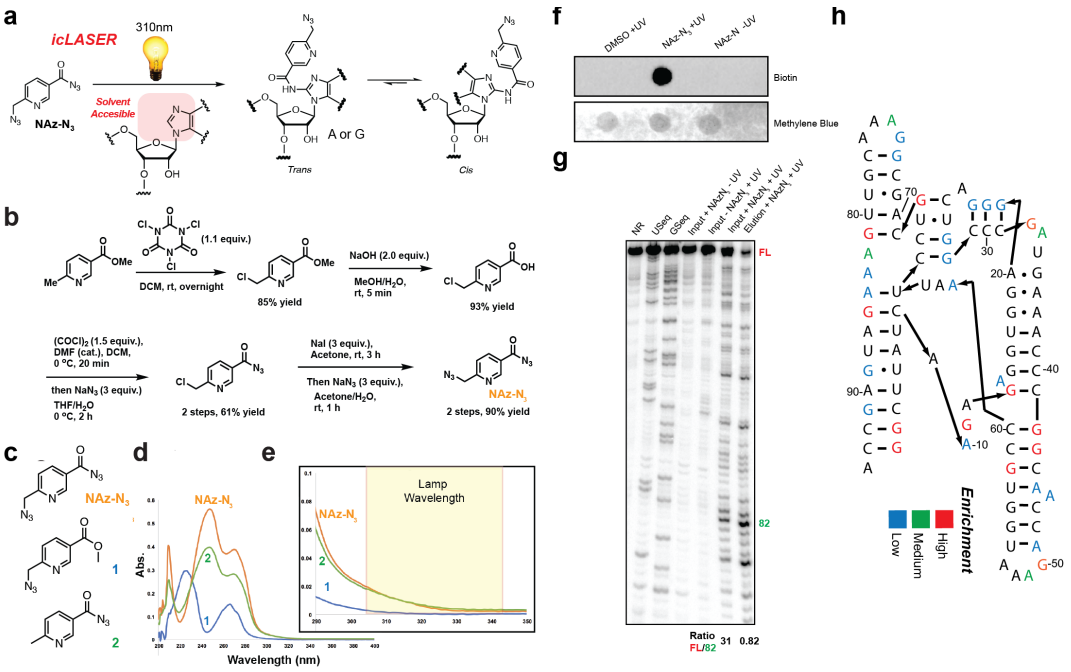


Figure 2: Development of a bi-functional probe for icLASER. a. Reaction schematic for LASER, with C-8 amidation utilizing NAz. **b.** Synthetic scheme for the probe NAz-N_3. **c.** Structure of NAz-N_3 and control probes for analysis of UV-VIS spectroscopy and light activation. **d.** UV-VIS spectra of compounds from panel c. **e.** Zoomed in UV-Vis spectra and corresponding wavelength used for light-activation of NAz-N_3 and control compounds in panel c. **f.** Streptavin-HRP dot blot of NAz-N_3 – modified RNA. RNA was incubated with NAz-N_3 in the presence of long-wavelength UV light. RNA was precipitated and conjugated with biotin using SPAAC, as denoted in Fig. 1. Followng SPAAC, RNA was blotted. **g.** Denaturing gel electrophoresis of modified SAM-I control RNA with NAz or NAz-N_3. RNA was incubated with NAz-N_3 in the presence of long-wavelength UV light. RNA was precipitated and conjugated with biotin using SPAAC. RNA was then enriched over magnetic streptavidin beads and eluted. Eluted RNA was reverse transcribed with ^{32}P labeled primer and cDNA analyzed on denaturing gel. Ratio of full length cDNA to modification was calculated against position 82 in the denaturing gel to demonstrate de-enrichment of the full length in comparison to the enriched modified position. **h.** Analysis of enriched positions from panel f.

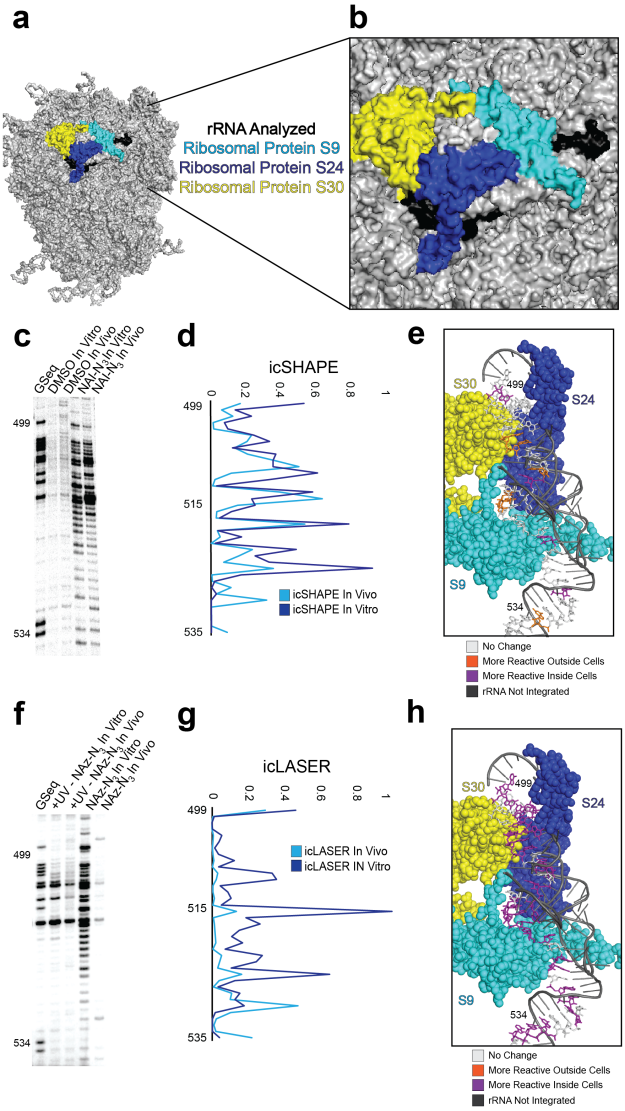
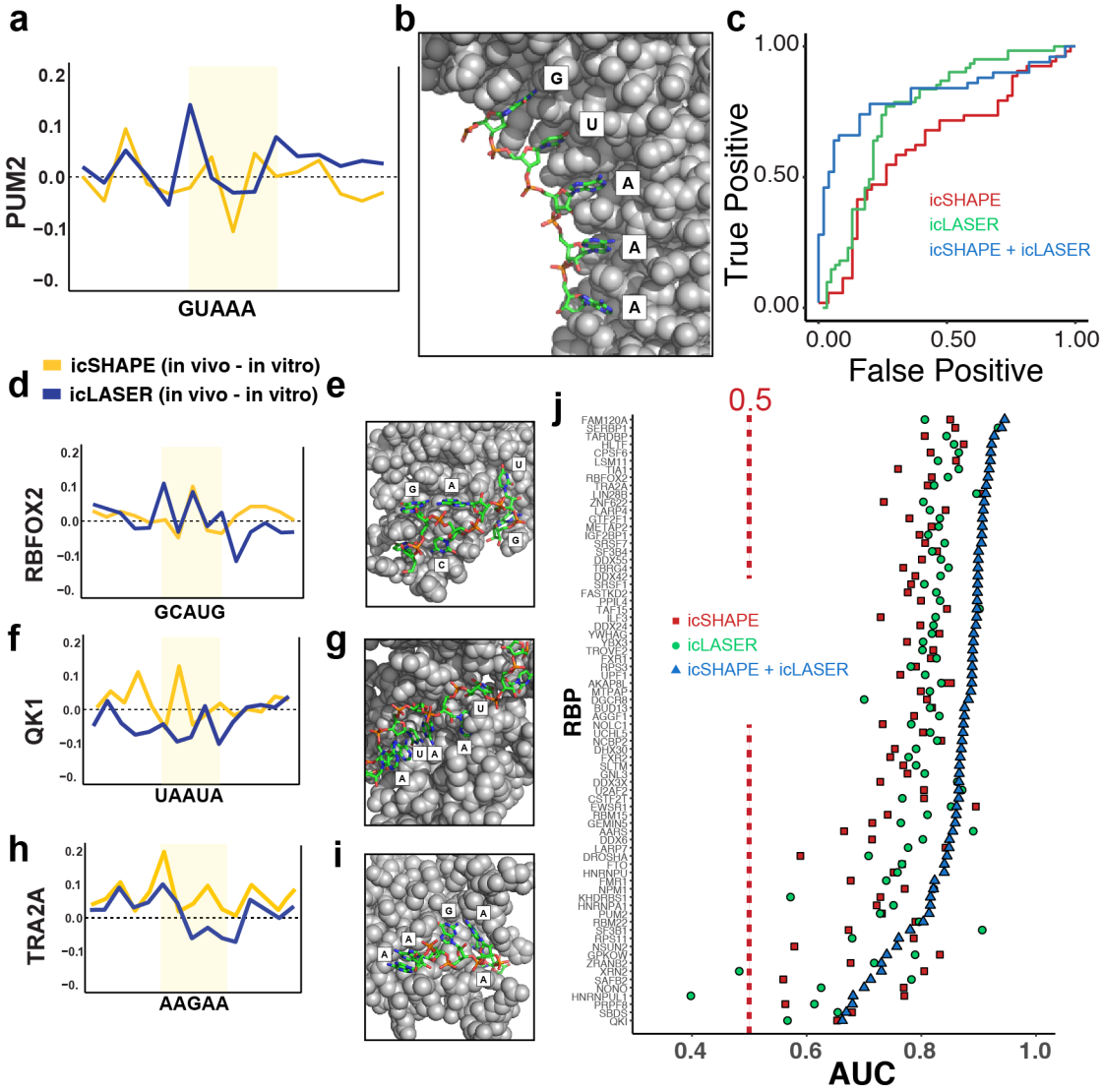


Figure 3: Analysis of icLASER and icSHAPE RT-stops. a. CryoEM model of the 80S ribosome (PDB 4v6x). **b.** Zoom in of rRNA section interrogated in for comparison footprinting. **c.** Denaturing gel analysis of NAI-N_3 reactivity. In vitro re-folded RNA or cells were subjected to incubation with NAI-N_3. Modification sites were analyzed by reverse transcription. **d.** icSHAPE reactivity profile over the same 18S rRNA region as in panel c. **e.** Structural analysis of 18S rRNA with differential reactivity from icSHAPE colored. **f.** Denaturing gel analysis of NAz-N_3 reactivity. In vitro re-folded RNA or cells were subjected to incubation with NAz-N_3. Modification sites were analyzed by reverse transcription. **g.** icLASER reactivity profile over the same 18S rRNA region as in panel f. **h.** Structural analysis of 18SrRNA with differential reactivity from icLASER colored.

533



534

535

536 **Figure 4: Utilizing icSHAPE and icLASER data to predict RNA-protein interactions**
537 **transcriptome-wide. a.** icSHAPE and icLASER difference maps between RNA probed inside
538 living cells and in vitro centered at the motif for PUM2. **b.** X-Ray structure of PUM2 bound to RNA
539 (PDB 3Q0Q). **c.** ROC analysis for predicting RNA-protein interactions for PUM2 eCLIP sites. **d.**
540 icSHAPE and icLASER difference maps between RNA probed inside living cells and in vitro
541 centered at the motif for RBFOX2. **e.** X-Ray structure of RBFOX2 bound to RNA (PDB 2ERR). **f.**
542 icSHAPE and icLASER difference maps between RNA probed inside living cells and in vitro

543 centered at the motif for QK1. **g.** X-Ray structure of QK1 bound to RNA (PDB 4JVH). **h.** icSHAPE
544 and icLASER difference maps between RNA probed inside living cells and in vitro centered at the
545 motif for TRA2A. **i.** X-Ray structure of TRA2A bound to RNA (PDB 2KXN). **j.** ROC analysis for
546 predicting RNA-protein interactions using icLASER and/or icSHAPE structure probing. For each
547 RNA-binding protein, we selected eCLIP bound sites *in vivo* and *in vitro*. A portion of this dataset
548 was used as a training set, and the remainder was used to test the classifier. The classifier was
549 trained using icSHAPE profiles, icLASER profiles, or both.
550

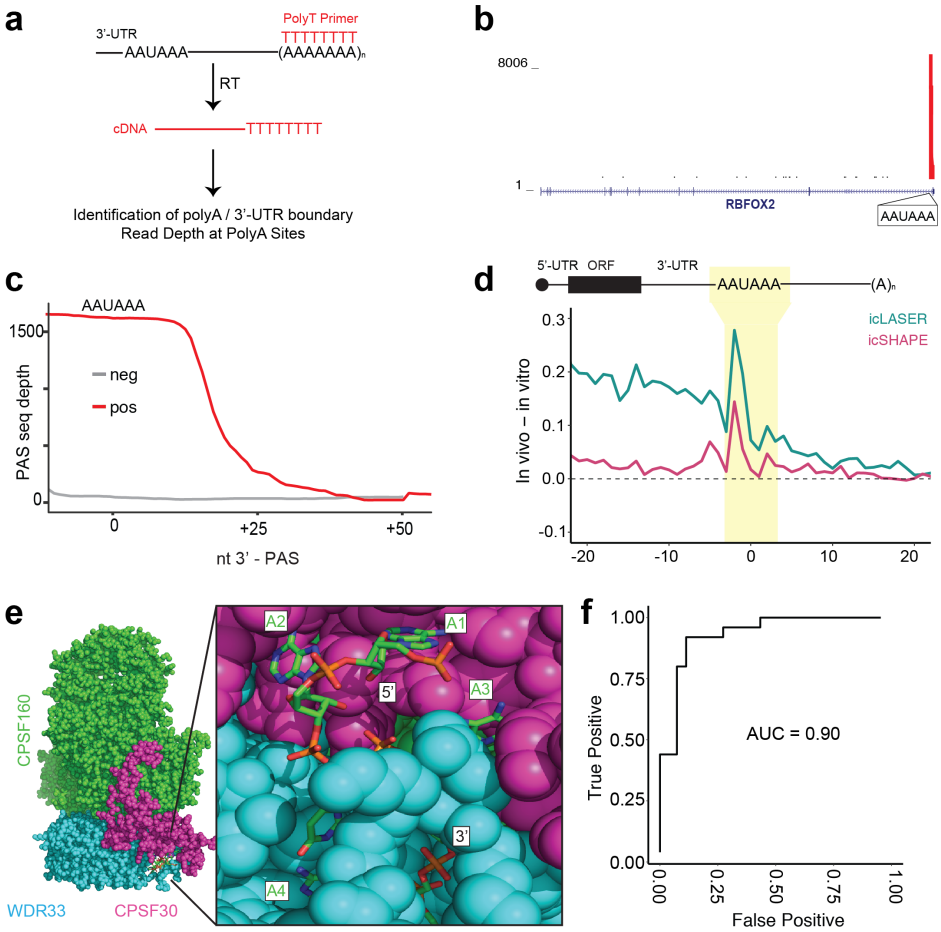


Figure 5: Utilizing icSHAPE and icLASER data to predict RNA polyA sites transcriptome-wide. a. Schematic of polyadenylation sequencing (PASseq). **b.** Genome browser track showing read density near the 3'-end of an RNA (RBFOX2) and demonstrating PAS seq specificity. **c.** Cumulative read density at PASseq-determined *in vivo* and *in vitro* polyA sites. The zero point on the X-axis is centered on the uridine residue of the AAUAAA motif. **d.** icSHAPE and icLASER data comparing in vivo and in vitro RNA structure profiling at the PAS site. **e.** Structure of the cryo-electron microscopy structure of a quaternary complex of human CPSF-160, WDR33, CPSF-30, and an AAUAAA RNA (PDB 6BLL). **f.** ROC analysis for predicting polyA sites using structure probing. The classifier was trained using icSHAPE and icLASER profiles.

563

564 **References**.

565 1. S. L. Klemm, Z. Shipony, W. J. Greenleaf, Chromatin accessibility and the regulatory
566 epigenome. *Nat Rev Genet* 10.1038/s41576-018-0089-8 (2019).
567 2. V. I. Risca, W. J. Greenleaf, Unraveling the 3D genome: genomics tools for multiscale
568 exploration. *Trends Genet* **31**, 357-372 (2015).
569 3. M. Kubota, C. Tran, R. C. Spitale, Progress and challenges for chemical probing of RNA
570 structure inside living cells. *Nat Chem Biol* **11**, 933-941 (2015).
571 4. E. J. Strobel, A. M. Yu, J. B. Lucks, High-throughput determination of RNA structures.
572 *Nat. Rev. Genet.* **19**, 615-634 (2018).
573 5. S. A. Mortimer, M. A. Kidwell, J. A. Doudna, Insights into RNA structure and function
574 from genome-wide studies. *Nat. Rev. Genet.* **15**, 469-479 (2014).
575 6. P. Tijerina, S. Mohr, R. Russell, DMS footprinting of structured RNAs and RNA-protein
576 complexes. *Nat. Protoc.* **2**, 2608-2623 (2007).
577 7. D. Mitchell, 3rd *et al.*, Glyoxals as in vivo RNA structural probes of guanine base-pairing.
578 *Rna* **24**, 114-124 (2018).
579 8. E. J. Merino, K. A. Wilkinson, J. L. Coughlan, K. M. Weeks, RNA structure analysis at
580 single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension
581 (SHAPE). *J. Am. Chem. Soc.* **127**, 4223-4231 (2005).
582 9. R. C. Spitale *et al.*, RNA SHAPE analysis in living cells. *Nat Chem Biol* **9**, 18-20 (2013).
583 10. R. C. Spitale *et al.*, Structural imprints in vivo decode RNA regulatory mechanisms.
584 *Nature* **519**, 486-+ (2015).
585 11. L. D. Poulsen, L. J. Kielpinski, S. R. Salama, A. Krogh, J. Vinther, SHAPE Selection
586 (SHAPES) enrich for RNA structure signal in SHAPE sequencing-based probing data.
587 *RNA* **21**, 1042-1052 (2015).
588 12. T. Adilakshmi, R. A. Lease, S. A. Woodson, Hydroxyl radical footprinting in vivo:
589 mapping macromolecular structures with synchrotron radiation. *Nucleic Acids Res.* **34**,
590 e64 (2006).
591 13. C. Feng *et al.*, Light-activated chemical probing of nucleobase solvent accessibility
592 inside cells. *Nat Chem Biol* **14**, 325 (2018).
593 14. V. Tarwade, O. Dmitrenko, R. D. Bach, J. M. Fox, The Curtius rearrangement of
594 cyclopropyl and cyclopropenoyl azides. A combined theoretical and experimental
595 mechanistic study. *J. Org. Chem.* **73**, 8189-8197 (2008).
596 15. M. Albuszis, P. J. Roth, W. Pauer, H. U. Moritz, Two in one: use of azide functionality for
597 controlled photo-crosslinking and click-modification of polymer microspheres. *Polymer*
598 *Chemistry* **7**, 5414-5425 (2016).
599 16. L. B. Xu, J. Farrell, R. G. Karunakaran, A. Honglawan, S. Yang, Synthesis of dual-
600 functional copolymer with orthogonally photosensitive groups. *Journal of Polymer*
601 *Science Part a-Polymer Chemistry* **51**, 1215-1222 (2013).
602 17. C. Feng *et al.*, Light-activated chemical probing of nucleobase solvent accessibility
603 inside cells. *Nat Chem Biol* **14**, 276-283 (2018).
604 18. R. C. Spitale *et al.*, Structural imprints in vivo decode RNA regulatory mechanisms (vol
605 519, pg 486, 2015). *Nature* **527**, 264-264 (2015).
606 19. M. M. Abdelsayed *et al.*, Multiplex Aptamer Discovery through Apta-Seq and Its
607 Application to ATP Aptamers Derived from Human-Genomic SELEX. *ACS Chem Biol*
608 **12**, 2149-2156 (2017).
609 20. Y. Ding *et al.*, In vivo genome-wide profiling of RNA secondary structure reveals novel
610 regulatory features. *Nature* **505**, 696-700 (2014).

611 21. R. C. Spitale *et al.*, Structural imprints in vivo decode RNA regulatory mechanisms.
612 *Nature* **519**, 486-490 (2015).
613 22. J. Boucas, Integration of ENCODE RNAseq and eCLIP Data Sets. *Methods Mol Biol*
614 **1720**, 111-129 (2018).
615 23. E. L. Van Nostrand *et al.*, Robust transcriptome-wide discovery of RNA-binding protein
616 binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508-514 (2016).
617 24. D. D'Amico *et al.*, The RNA-Binding Protein PUM2 Impairs Mitochondrial Dynamics and
618 Mitophagy During Aging. *Mol Cell* **73**, 775-787 e710 (2019).
619 25. N. J. Proudfoot, Ending the message: poly(A) signals then and now. *Genes Dev.* **25**,
620 1770-1782 (2011).
621 26. P. J. Shepard *et al.*, Complex and dynamic landscape of RNA polyadenylation revealed
622 by PAS-Seq. *Rna-a Publication of the Rna Society* **17**, 761-772 (2011).
623 27. J. Harrow *et al.*, GENCODE: The reference human genome annotation for The
624 ENCODE Project. *Genome Res.* **22**, 1760-1774 (2012).
625 28. Y. D. Sun *et al.*, Molecular basis for the recognition of the human AAUAAA
626 polyadenylation signal. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E1419-E1428 (2018).
627 29. J. Brown, M. Pirrung, L. A. McCue, FQC Dashboard: integrates FastQC results into a
628 web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* **33**,
629 3137-3139 (2017).
630 30. D. R. Zerbino *et al.*, Ensembl 2018. *Nucleic Acids Res* **46**, D754-D761 (2018).
631 31. C. A. Sloan *et al.*, ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**, D726-
632 732 (2016).
633