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 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.

37 **Introduction**

38 Precise structural conformations are a hallmark of functional nucleic acid polymers in cells. For
39 example, the genome is arranged in a compact three-dimensional structure, whose dynamics and
40 solvent accessibility are critical for the control of gene expression(1). Methods to measure these
41 structural properties of the genome are now quite mature and can be done genome-wide to
42 understand the interactions between DNA and proteins and even predict RNA expression
43 outcomes(2). As a corollary, there has also been a more recent emergence of approaches to
44 measure transcriptome-wide RNA structure in cells(3), but these investigations have had limited
45 utility in understanding RNA interactions with trans factors. Further, while many complementary
46 structure probes have been developed, their use in combination to better understand complex
47 RNA structures and RNA-protein binding has been limited.

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 be used to identify base-paired residues within RNA. For example,
51 adenosine and cytidine residues not involved in Watson-Crick-Franklin pairing can be alkylated
52 by 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, often uses 2'-hydroxyl flexibilities as a proxy for single-strandedness of a
55 residue(8, 9). However, unique aspects of RNA structure that extend beyond base-pairing or
56 base-flexibility. provide valuable information in measuring RNA structure and function.

57 Developing bifunctional RNA structure probing reagents that enable reactivity with RNA
58 and enrichment of reaction sites has been demonstrated to dramatically increase signal-to-noise
59 ratios of RT-stops arising from chemical adducts(10, 11). We previously developed icSHAPE,
60 which features a bifunctional chemical probe to measure nucleobase flexibility through RNA
61 hydroxyl acylation transcriptome wide(10) (**Fig. 1, a**). Solvent accessibility, an orthogonal
62 property of RNA, has been more difficult to probe inside cells, requiring the use of a synchrotron

63 radiation light source(12). We recently described the development of LASER, or Light Activated
64 Structural Examination of RNA(13). LASER relies on the facile light activation of a nicotinoyl azide
65 to a nitrenium ion, which undergoes electrophilic aromatic substitution with electron rich purine
66 residues. While this initial work demonstrated a strong correlation between purine C-8 solvent
67 accessibility to LASER reactivity. Here we designed, synthesized, and implemented a bifunctional
68 LASER probe to measure transcriptome-wide purine C-8 solvent accessibility (**Fig. 1, b,** *in vivo*
69 click LASER, icLASER). Our design employed similar azide functional groups for enrichment to
70 allow for a common biochemical and computational approach to identify RNA-reagent adducts for
71 measuring RNA structure (**Fig. 1, c**).

72 By implementing this novel bi-functional probe, icLASER provides the first transcriptome-
73 wide measurement of solvent accesibility. Towards the goal of integrating multiple, orthogonal
74 measurements of RNA structure, we develop an approach to directly compare icSHAPE (hydroxyl
75 acylation; flexibility) and icLASER (solvent accessibility). We highlight the power these data and
76 this strategy in the context of predicting RNA-protein interactions and RNA polyadenylation. Our
77 results demonstrate that combinatorial RNA structure probing can be employed to compliment
78 orthogonal structure probing methods to better understand RNA structure and processing in cells
79 transcriptome-wide.

80

81 **Results**

82 **Development of a bifunctional LASER probe**

83 An optimal probe for icLASER would retain the selectivity of NAz and also allow for
84 enrichment using conditions non-harmful to RNA. As such, NAz-N_3, which we predicted would be
85 able to undergo light transformation to the activated nitrenium ion, would also preserve the alkyl
86 azide for Cu-free “click” after RNA adduct formation. Notably, aroyl azides are amenable to
87 activation by lower energy long wavelength light, whereas alkyl azides need higher energy short

88 wavelength light for activation(14-16). We hypothesized these special characteristics of azide
89 stability would permit photo-specific activation of NAz-N_3 for icLASER probing.

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

99 We obtained UV-Vis spectra of these compounds and observed that only NAz-N_3 and **2**
100 have absorption in the long wavelength UV region used previously by us in LASER and known to
101 be specific for aroyl azides (**Fig.2 , d&e**)(13). Consistent with what others have observed with
102 azide-specific activation, these data suggest that light activation of NAz-N_3 should be specific to
103 the aroyl azide. To test reactivity and preservation of a SPAAC-compatible azide, we incubated
104 an *in vitro* transcribed SAM-I riboswitch with NAz-N_3 and exposed the solution to (+/-) long
105 wavelength UV light. Following irradiation, we isolated the RNA and incubated the RNA with
106 dibenzylcyclooctyne-conjugated to biotin (DBCO-biotin). We observed biotin conjugation on a
107 streptavidin dot blot only in the sample incubated with long wavelength UV light and NAz-N_3
108 (**Fig.2, f**).

109 We next tested if NAz-N_3 enriched RNA would produce similar truncation profiles to NAz
110 probing. We used the biotinylated NAz and NAz-N_3 modified riboswitch RNA and subjected them
111 to streptavidin-coated magnetic bead enrichment. For each experiment, we compared three
112 distinct populations of RNA. A portion of the input RNA was set aside as a control, and we
113 subsequently isolated both the enriched and flow through sample. Each sample was then reverse

114 transcribed with a ^{32}P-cDNA primer and analyzed by denaturing gel electrophoresis. As shown in
115 **Fig2. g, h**, the NAz and NAz-N_3 input samples both showed indistinguishable structure cDNA
116 patterns. NAz flow through contained RT-stops and significant amount of full length (FT), with no
117 detectable cDNA molecules in the elution. In stark contrast, NAz-N_3 flow through cDNAs were
118 almost all full length and the elution contained significantly higher signal for structure stop cDNA
119 and dramatically reduced full-length band. Notably, full-length bands in NAz-N_3 are modifications
120 near the 5'-end of the RNA and not background cDNAs. Finally, NAz-N_3 elution mapped onto sites
121 of modification similar observed with NAz (A & G residues; **Fig. 2, h**). Further, we also
122 demonstrated that NAz-N_3 can modify RNA and measure RNA structure in cells, and observed
123 differences *in vitro* and in cells that correspond to cDNA profiles to that of the parent compound
124 NAz(17) (**Supplementary Figure 1**). Overall, this result demonstrates that NAz-N_3 is capable of
125 enrichment using SPAAC reactions, yet still yields consistent truncation signatures for structural
126 analysis.

127 **Transcriptome-wide implementation of NAz-N_3**

128 We modified K562 cells (*in vivo*) and K562 total RNA isolated and re-folded outside of cells
129 (*in vitro*) in independent experiments with NAz-N_3 (icLASER) and NAI-N_3 (icSHAPE). We applied
130 our well-established protocol for the isolation of poly-adenylated (polyA) RNA and mapping of
131 biotin-conjugated RNA structure probes through deep sequencing (**Figure 1, Supplementary**
132 **Methods)** (18, 19), which generated more than 100 million reads for each dataset
133 (**Supplementary Figure 2**).

134 Mapping RT stops from icLASER enriched libraries revealed a marked enrichment of A
135 and G residues, which was expected given the reactivity profile of the NAz-N_3 probe
136 (**Supplementary Figure 3**). This profile was unique to icLASER, whereas for icSHAPE we
137 observed a slight enrichment for adenosine stops, as seen previously (**Supplementary Figure**
138 **3**) (10). Importantly, we find a strong correlation between +UV samples, whereas, the distribution
139 of +UV and –UV RT-stops showed very low correlation (**Supplementary Figure 4**). Comparison

140 of the icSHAPE and icLASER stops showed strong intra-reagent correlation but weaker inter-
141 reagent correlation, suggesting each probe measured a discrete set of positions in the
142 transcriptome.

143 As an initial quality control, we analyzed the icLASER and icSHAPE data against manual
144 footprinting on an RNA structure in which we had reasonably high-resolution data: the ribosome.
145 We focused on a section of 18S rRNA has been previously interrogated by footprinting(13) (**Fig.**
146 **3, a & b**). The manual footprinting for NAI-N_3 in this region showed some slight differences
147 between in and outside of cells (**Fig. 3, c**). In similar fashion, our icSHAPE data had similar overall
148 reactivity between the two experimental setups (**Fig. 3, d**). Analysis of the RNA in this region (Fig.
149 3, e) revealed that it was solvent protected by rRNA and ribosome binding proteins, but was
150 largely still single stranded and thus would be reactive toward SHAPE reagents. We also
151 performed comparison probing with NAz-N_3 on the same stretch of rRNA and observed marked
152 differences between in and outside the cells (**Fig. 3, f**). Consistent with this analysis was the
153 reactivity profile for icLASER (**Fig. 3, g**), which is supported by the solvent protected nature of
154 this stretch of RNA in the 18S ribosome CryoEM model (**Fig. 3, h**). Overall, these results suggest
155 that icLASER and icSHAPE are both measuring environment-unique aspects of RNA structure in
156 and outside cells.

157 As this is the first transcriptome-wide map of solvent accessibility probing with icLASER
158 we also analyzed mRNA profiles. We observed that start codons are largely open and solvent
159 accessible with high reactivity at the A and G of the AUG (**Supplementary Figure 5**). The stop
160 codon also displayed high reactivity in the last two positions (**Supplementary Figure 5**), and both
161 of these observations are in line with previous measurements using icSHAPE and DMS-seq
162 (20).To understand what regions of icLASER and icSHAPE signal could be due to conditions
163 specific to living cells, we compared the *in vivo* and *in vitro* profiles as previously described for
164 icSHAPE (VTD, (21)). Specifically, we examined all hexamers across the transcriptome, and
165 consistent with our previous reports, icSHAPE structure was overall higher *in vivo*

166 (**Supplementary Figure 6.**). This is thought to be correlated with structural remodeling *in vivo*, in
167 comparison to *in vitro*(10). In contrast, we observed a bimodal distribution of solvent accessibility
168 differences in and outside of cells with icLASER. Hexamers that include the Kozak sequence
169 showed small VTD differences, which is in line with previous reports with icSHAPE(21). However,
170 for icLASER, hexamers containing RNA binding protein motifs had lower average VTD values.
171 This observation is consistent with the known reactivity of LASER probing, which can result in a
172 change in chemical reactivity in cells due to protein footprinting, suggesting that icLASER could
173 be used to map protein-RNA interactions. Overall, these results demonstrate the robustness of
174 icLASER probing and demonstrate its utility in mapping the structures of RNAs in and outside of
175 cells.
176
177 **Global prediction of protein footprints by combining structure probing tools.**
178 Despite the importance of RNA-protein interactions and the interplay of RNA structure and
179 protein binding, there has thus far been a lack of transcriptome-wide analysis of RNA structure
180 motifs with the express goal of attempting to predict binding sites. To address this gap, we
181 assessed if icLASER, icSHAPE, or their combination could be used to predict RNA-protein
182 interactions (**Fig. 4**). As a reference for experimentally measured RNA binding protein (RBP) sites,
183 we leveraged enhanced individual nucleotide resolution CLIP (eCLIP) datasets from ENCODE
184 collected from K562 cells (22, 23). We focused on RBPs that had significant binding in 3'-UTRs
185 (ensuring high sequence coverage in icLASER and icSHAPE data), binding motifs with at least
186 three A or G residues, targets expressed at greater than or equal to 5 RPKM. As control regions,
187 we sampled the same number of sites with matching RNA motifs not called as peaks in the eCLIP
188 data.
189 We first focused on PUM2, a well-established RNA binding protein that is important in
190 regulating RNA localization and translation, and has high-resolution structures for comparison
191 (**Fig. 4, a**)(24). Comparative icLASER and icSHAPE profiles in and outside of cells yielded key

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

208 We expanded our initial analysis to other RBPs that have reported crystal or NMR
209 structures bound to cognate RNA sequences. RBFOX2 (**Fig. 4, d&e**) displayed a similar icSHAPE
210 structure profile to our previous efforts with icSHAPE in mouse embryonic stem cells(21).
211 icLASER profile differences showed high reactivity at the G1 position, and correspondingly the C-
212 8 of G1 is pointed out to solvent in the crystal structure. The same is true for A3, where the C-8
213 position is solvent exposed. Quaking1 (QK1) (**Fig. 4, f&g**) shows generally lower icSHAPE
214 reactivity across its binding motif. These positions are also protected from solvent either through
215 direct protein interactions or shielding from phosphate backbone, as is the case for A3 in the motif.
216 Lastly, TRA2A (**Fig. 4, h&i**) displayed correlation at positions 1 and 2 (AA); these are both solvent
217 exposed and likely to be highly dynamic, and as such had high icLASER and icSHAPE reactivity.

218 The last two A residues were protected from solvent by protein interactions and would likely be
219 stabilized to limit their SHAPE reactivity. Overall, these data show icLASER and icSHAPE provide
220 non-overlapping nucleotide resolution at RBP binding sites that can read out the biophysical
221 conformation of the bound RNA in a manner that corresponds to each method's chemical
222 properties.

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

236

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

238 Our analysis above (**Fig. 3d**) revealed hexameric sequences that have differences inside
239 and outside of cells. One such sequence that had higher in-cell reactivity for icLASER (and
240 icSHAPE) was the polyadenylation signal sequence (PAS) within the 3'-UTRs(25). Given the
241 measurable VTD value, we explored if our structure probing data could be used predict
242 polyadenylation transcriptome-wide. This would be an extremely important further validation of

243 our approach to use structure probing for predicting RNA-protein interactions and potential
244 processing events.

245 To experimentally determine PAS sites, for SVM analysis, we generated the first
246 polyadenylation sequencing data (PAS-seq (26)) for K562 cells. PAS-seq uses polyA tail priming
247 to identify the sites of polyA tail selection (**Fig. 5, a**)**.** Inspection of PAS reads demonstrated clear
248 read buildup at the 3'-end of transcripts (**Fig. 5, b**) and we obtained sites of high and low PAS
249 read depth (**Fig. 5, c**). We then compared the icLASER and icSHAPE profiles at the PAS sites
250 and sequences with the same motif which are not annotated to be PAS sequences (negative
251 control) (27). We noticed a striking structural difference between them: a large peak at the first
252 two adenosine residues, followed by a drop in icLASER signal for the remaining UAAA (**Fig. 5,**
253 **d**). To understand if these differences were related to a biophysical conformation of the RNA in
254 an active PAS-recognition complex, we examined a newly published structure(28). This structure
255 contains three large proteins which form into a complex and recognize the AAUAAA motif (**Fig. 5,**
256 **e**). Close inspection of the structure revealed that the two five prime adenosine residues are
257 completely solvent exposed, whereas the UAAA residues are solvent protected by protein binding.
258 Impressively, the two AA residues have their C-8 positions (site of icLASER reaction) completely
259 exposed (**Fig. 5, e**). Consistent with our icLASER data, sites with high icLASER signal at the first
260 two adenosine residues had very high coverage by PAS-seq. This data further suggested to us
261 that icLASER signal alone could be used to predict polyA site selection. To test this hypothesis,
262 we utilized SVM(10) and demonstrated that icLASER and icSHAPE used together had an AUC
263 of 0.9 for predicting PAS sites (**Fig. 5, f**). These data nicely demonstrate that icLASER (and
264 combined icLASER and icSHAPE) can be used to predict sites of posttranscriptional regulation
265 and could be integrated with orthogonal datasets to interpret posttranscriptional processing of
266 RNAs.
267
268

269 **Discussion**

270 Here, by developing a bifunctional enrichment handle, we extend the utility and flexibility
271 of our previously reported RNA solvent accessibility probe, NAz. The new reagent, NAz-N_3,
272 similarly relies on specific photoactivation of an aroyl azide by long-wavelength UV light, but can
273 subsequently be ligated to biotin using copper-free "click" chemistry, for enrichment of modified
274 sites of adduct formation. Using this chemistry, we measured RNA solvent accessibility in K562
275 cells, transcriptome-wide.

276 With the development of transcriptome-wide RNA structure probing techniques, an
277 exciting but thus far poorly explored aspect of these data has been the possibility to infer
278 interactions between RNAs and their cellular partners. By employing a computational strategy
279 (SVM) to combine icSHAPE and icLASER (as well as other reactivity-based measurements) we
280 take a critical step towards learning the potential of these methods in predicting RBP binding and
281 other functional RNA processing activities like PAS selection. We demonstrate the power of this
282 approach by predicting RNA-protein binding sites for a large number of RNA binding proteins. We
283 demonstrate that such an approach could be very powerful for measuring RNA-protein interfaces
284 and is highly complementary with protein-centric methods such as eCLIP. Further, we utilize RNA
285 structure probing to identify a structure signature associated with polyadenylation sites, which is
286 due to the presence of a protein bound at the PAS site. This extension also enables structure
287 probing to be utilized for other aspects of RNA biology, such as RNA processing.

288 The complexity of nucleic acid polymer structure is now well appreciated for DNA and the
289 folding of the genome; importantly, measuring the structure of such nucleic acids and how proteins
290 interact with DNA has been incredibly valuable for understanding how the genome is regulated to
291 control biological processes within cells. Our goal herein is to make headway toward this goal for
292 RNA, by the continued development of tools for transcriptome-wide measurement of RNA
293 structure and RNA interactions that can contribute to its biological function and regulation. We
294 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.

326 **Primers used for Reverse Transcription:**

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

328 18s Primer 1 (5', CCAATTACAGGGCCTCGAAA, 3')

329

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

331 A 94 nucleotide construct consisting of the sequence for the SAM-I riboswitch from the metF-
332 metH2 operon of *T. tencongensis* was designed into a plasmid with IDT. The plasmid encoding
333 the SAM-I sequence was transformed into a One Shot Top 10 chemically competent cells
334 (ThermoFisher) and plated on lysogeny broth (LB, ThermoFischer) supplemented with 100
335 mg/mL ampicillin (VWR) agar plates. A single colony was selected in a 3 mL culture and grown
336 overnight. The resulting plasmid was isolated using a QIAprep Miniprep kit (Qiagen). Transcription
337 template was prepared by PCR in 50 μL volumes using primers directed against the T7 promoter
338 (5', TAATACGACTCACTATAGGG, 3') and an adaptor sequence for reverse transcription (5',
339 ATTTAGGTGACACTATAGTT, 3') with Phusion High-Fidelity PCR master mix (NEB). The
340 mixture was treated with 1 μL of DPN1 (NEB) restriction enzyme and incubated at 37 °C for 30
341 min. The resulting PCR product was purified using a QIAquick PCR purification kit (Qiagen). *In*
342 *vitro* transcription of SAM-I RNA was performed in 20 μL reactions using 800 ng of template SAM-
343 I DNA with a T7 Ribomax transcription kit (Promega). Samples were treated with 1 μL of RQ1
344 DNAse (Promega) at 37 °C for 30 min. The resulting solution was brought up to 100 μL and
345 precipitated with 10 μL of 3 M NaOAc (pH 5.2), 1 μL glycoblue (ThermoFisher), and 300 μL
346 ethanol. The resulting RNA stored at -80 °C for 30 minutes, centrifuged at 15,000 rpm for 15
347 minutes, and resuspended in 50 μL of RNase-free water. An aliquot of SAM-I was run on a
348 denaturing PAGE gel (15% polyacrylamide, 0.5X TBE, 7 M urea) alongside a 100 bp RNA ladder
349 (Invitrogen). The band of interest was visualized with 1x SYBR-gold stain (ThermoFisher) in water

350 for 15 minutes. Resulting concentrations of RNA was quantified by integrating intensity of the
351 ladder with the RNA band of interest.
352
353 **^{32}P End labeling for reverse transcription:**
354 Primer DNA was 5' end labeled in 10 μL reactions in a T4 PNK mix (1 μL 10x T4PNK reaction
355 buffer, 2 μL 100 mM DNA primer, 5 μL nuclease free water, 1 μL γ-^{32}P-ATP, 1 μL T4 PNK, NEB).
356 The reaction was allowed to proceed for two hours at 37 °C. Reactions were stopped by the
357 addition of 5 μL of Gel Loading Buffer II. The reaction was loaded onto a 15% denaturing PAGE
358 gel. The band of interest was visualized by a phosphorimager (Typhoon, GE healthcare). The
359 resulting band was excised and eluted overnight in 400 μL of 300 mM KCl. Resulting solution was
360 EtOH precipitated and dissolved to 8,000 counts per minute (cpm)/μL for further use in reverse
361 transcription.
362
363 **Modification of RNA *in vitro*:**
364 5 μg total RNA (isolated from HeLa cells) or 10 pmol of *in vitro* transcribed RNA in 6 μL metal-
365 free water was heated for 2 minutes at 95 °C. The RNA was then flash cooled on ice for 1 minute,
366 and brought to room temperature. 3 μL of 3x RNA folding buffer (333 mM HEPES, pH 8.0, 20 mM
367 $MgCl_2$ and 333 mM NaCl) was added, and the RNA was allowed to equilibrate at 37 °C for 5 min.
368 To this mixture 1 μL of 3 M NAz, 1 M NAz-N_3 in DMSO (+) or DMSO (-) was added. Reactions
369 were then exposed to 20 Watt lamp (Zilla Desert UVB 50) UV light for 3 minutes for NAz, and 10
370 minutes for NAz-N_3. Reactions were brought up to 200 μL water, and extracted once with 200 μL
371 acid phenol/chloroform/isoamyl alcohol (pH 4.5, Ambion), and washed twice with 200 μL
372 chloroform (Sigma). Samples were precipitated by adding 20 μL 3 M NaOAc (pH 5.2), 1 μL
373 glycoblue (20 μg/μL) and 600 μL EtOH. Pellets were washed twice with 70% cold ethanol and
374 resuspended in 5 μL nuclease free water.

375

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

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

384

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

386 Hybond N+ membranes (GE) were pre-incubated in 10X SSC. Precipitated biotinylated total RNA
387 was dissolved in 2 μL of RNase free water. RNA was loaded onto the Hybond membrane and
388 crosslinked using 254 nm ultraviolet light. The membrane was incubated with blocking solution
389 (120 mM NaCl, 16 mM Na_2HPO_4, 8 mM NaH_2PO_4, 170 mM SDS) for 30 min. To the membrane
390 was added 1 μL Pierce high sensitivity streptavidin-HRP(ThermoFisher) in blocking solution. The
391 membrane was washed twice with wash buffer A (1:10 blocking solution) for 30 min, and twice
392 with wash buffer B (100 mM Tris pH 9.5, 100 mM NaCl, 20 mM $MgCl_2$) for 5 min. Membrane was
393 incubated with Pierce Western blotting substrate and visualized on the ChemiDoc (Biorad) under
394 chemiluminesence hi sensitivity. To visualize RNA, the membrane was stained with a methylene
395 blue solution (0.2% w/v methylene blue, 0.4 M sodium acetate).

396

397 **Modification of RNA in cells:**

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

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

425

426

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

428 ^{32}P-end-labeled DNA primers were annealed to modified RNA by incubating 95 °C for two
429 minutes, then 25 °C for two minutes, and 4 °C for 2 minutes. To the reaction, 1 μL of 5x First
430 strand buffer, 0.5 μL nuclease free water, 0.5 μL 100 mM DTT, and 0.5 μL 10 mM dNTP's were
431 added. The reaction was preincubated at 52 °C for 1 min, then 1 μL superscript III (ThermoFisher)
432 was added. Extensions were performed for 15 min. To the reaction, 1 μL 4 M sodium hydroxide
433 was added and allowed to react for 5 min at 95 °C. The resulting complementary DNA (cDNA)
434 was snapped cooled on ice, and ethanol precipitated according to above procedures. Purified
435 cDNA was resuspended in 2 μL of nuclease-free water and 2 μL of Gel Loading Buffer II. cDNA
436 products were resolved on 10% denaturing polyacrylamide gel, and visualized by a gel imager
437 (Typhoon, GE healthcare).

438

439 **<u>Generation of PolyA sequencing libraries.</u>**

440 10 µg total RNA extracted with Trizol (Ambion) was fragmented with fragmentation reagent
441 (Ambion) at 70 °C for 10 minutes followed by precipitation with ethanol. Reverse transcription was
442 performed with PASSEQ7-2 RT oligo:

443

444 [phos]NNNNAGATCGGAAGAGCGTCGTGTTCGGATCCATTAGGATCCGAGACGTGTGCTCT
445 TCCGATCTTTTTTTTTTTTTTTTTTTTT[V-Q]

446

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

450 heated at 95 °C for 2 minutes, and then cool to 37 °C slowly. Circularized cDNA was linearized
451 by BamH I (Promega). cDNA was collected by centrifugation after ethanol precipitation. PCR was
452 carried out with primers PE1.0 and PE2.0 containing index (Illumina). Around 200 bp of PCR
453 products was gel-purified and submitted for sequencing (single read 100 nucleotides).
454
455 **Data analysis for icSHAPE and icLASER**
456 Raw sequence reads were quality-checked using FastQC (29) and demultiplexed. Reads were
457 aligned to the ENSEMBL Release 88 GRCh38 transcriptome (30)and per-position enrichment
458 scores were calculated using the icSHAPE pipeline(21). Default parameters were used with the
459 exception of the final filtering step, where minimum values for hit coverage and background base
460 density were removed (-T 0 -t 0).
461
462 Publicly available eCLIP peak data for five RNA binding proteins in K562 cells was downloaded
463 from the ENCODE project(31). Known 5bp binding motifs BiorXiv
464 https://doi.org/10.1101/179648)) for each protein were located in each peak; peaks lacking the
465 motif were discarded. icLASER enrichment scores were extracted for each motif position, plus
466 five basepairs up- and downstream from the motif. Negative control sites were identified as
467 occurrences of the same 5bp motif that fall outside of eCLIP peaks. As with binding sites,
468 icLASER enrichment scores were extracted for a 15bp range centered on the motif. A number of
469 negative sites equal to the number of positive sites was selected randomly from the pool of
470 possible negatives.
471
472 icLASER analysis at poly-A signal sequences was conducted similarly; however, Ensembl-
473 annotated polyadenylation signal sequences were used to identify positive sites, and enrichment
474 scores were retrieved for a range of 20bp up- and downstream of the 6bp polyadenylation signal
475 motif.

476

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

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

484

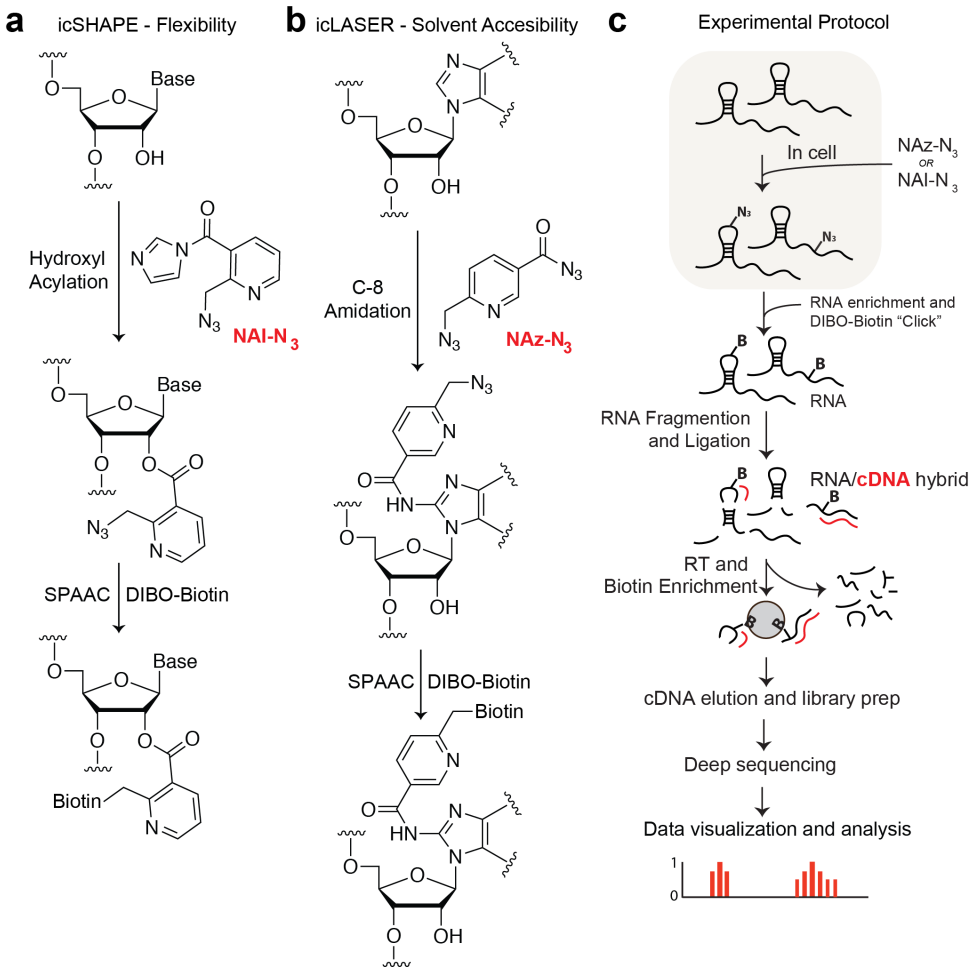
485 **SVM analysis of icSHAPE/icLASER and polyadenylation sites**

486 SVM analysis was conducted as with RNA-binding proteins. icSHAPE or icLASER scores for a
487 46bp range centered on the polyadenylation signal sequence were used as features.

488

489 **Figures.**

490



491

492

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

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

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

496 wide.

497

498

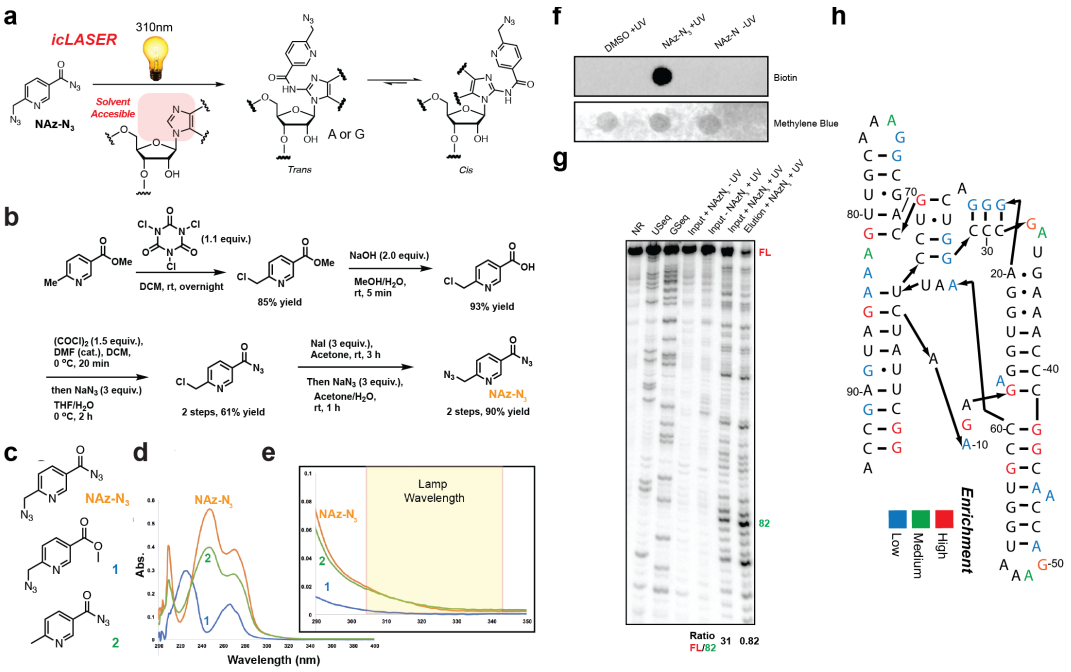
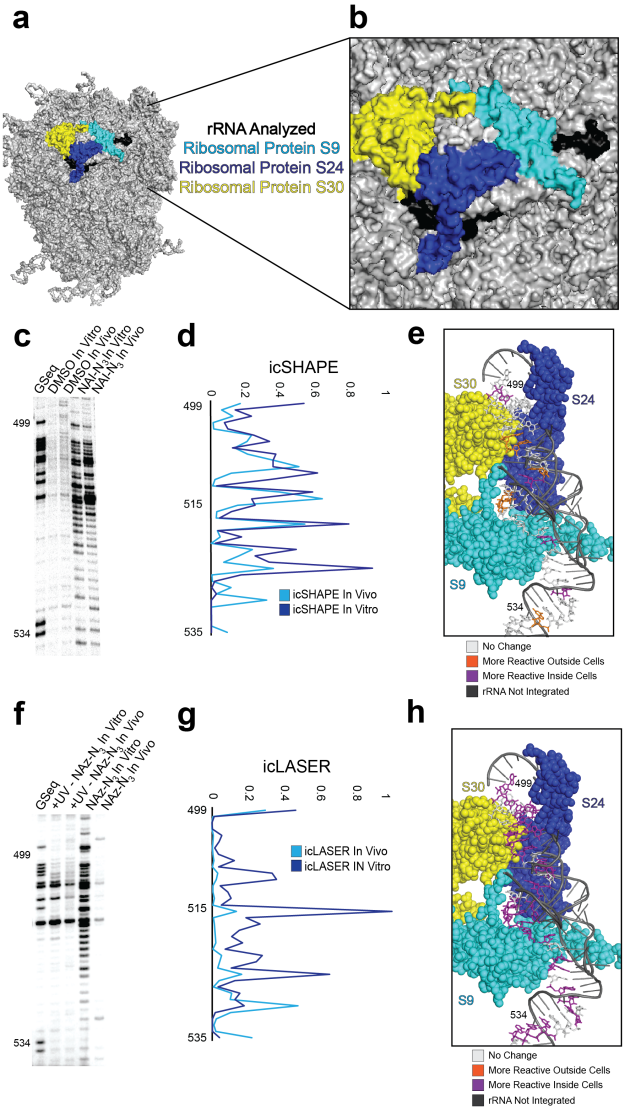


Figure 2: Development of a bi-functional probe for icLASER. a. Reaction schematic for LASER, with C-8 amidation utilizing NAz. **b.** Synthetic scheme for the probe NAz-N_3. **c.** Structure of NAz-N_3 and control probes for analysis of UV-VIS spectroscopy and light activation. **d.** UV-VIS spectra 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.



515

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

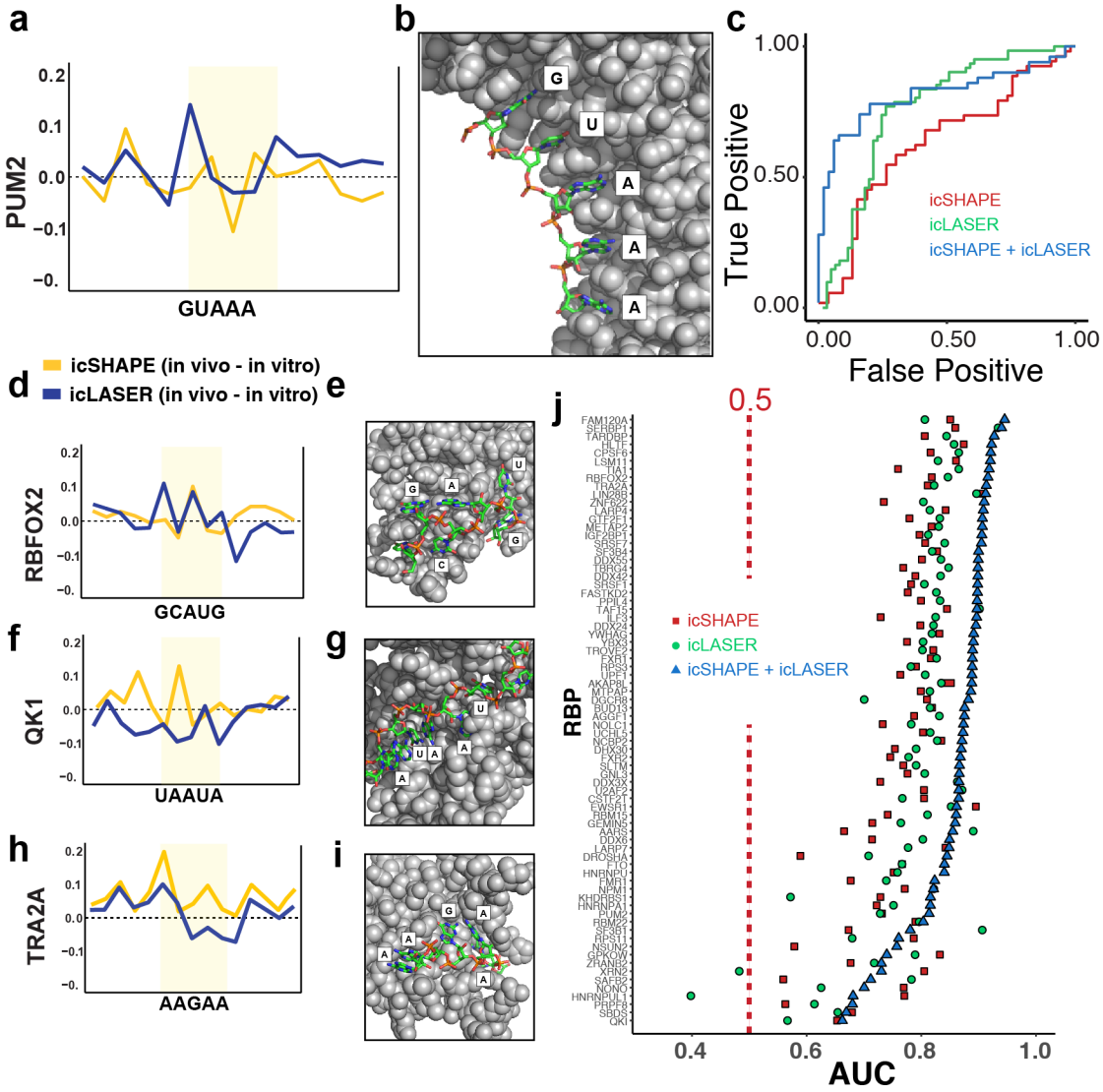


Figure 4: Utilizing icSHAPE and icLASER data to predict RNA-protein interactions transcriptome-wide. a. Schematic of methods employed to utilize icSHAPE and icLASER data to footprint RNA-protein interactions and utilize such footprints to predict protein occupancy. **b.** icSHAPE and icLASER difference maps between RNA probed inside living cells and in vitro centered at the motif for RBFOX2. **c.** X-Ray structure of RBFOX2 bound to RNA (PDB 2MGZ). **d.** icSHAPE and icLASER difference maps between RNA probed inside living cells and in vitro centered at the motif for PUM2. **e.** X-Ray structure of PUM2 bound to RNA (PDB 3Q0Q). **f.**

535 icSHAPE and icLASER difference maps between RNA probed inside living cells and in vitro
536 centered at the motif for QK1. **g.** X-Ray structure of QK1 bound to RNA (PDB 4JVH). **h.** icSHAPE
537 and icLASER difference maps between RNA probed inside living cells and in vitro centered at the
538 motif for TRA2A. **i.** X-Ray structure of TRA2A bound to RNA (PDB 2KXN). **j.** ROC analysis for
539 predicting RNA-protein interactions using icLASER and/or icSHAPE structure probing. For each
540 RNA-binding protein, we selected eCLIP bound sites *in vivo* and *in vitro*. Half of this dataset was
541 used as a training set, and the other half was used to test the classifier. The classifier was trained
542 using icSHAPE profiles, icLASER profiles, or both.
543

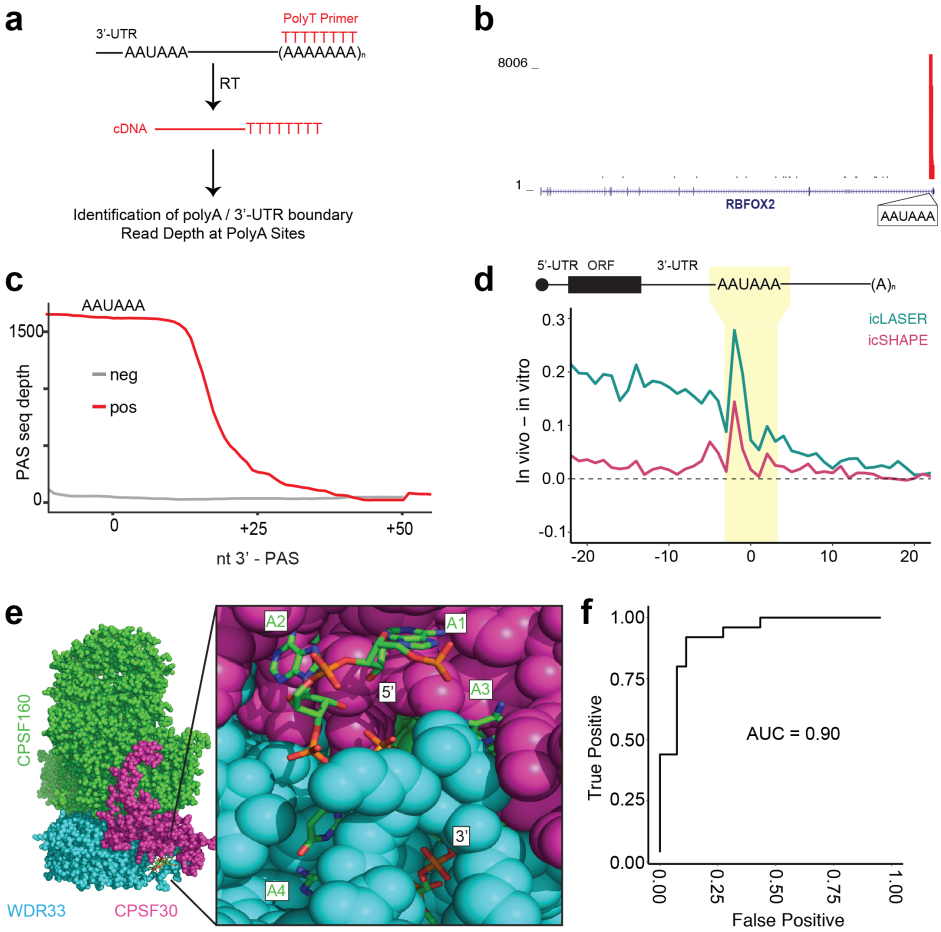


Figure 5: Utilizing icSHAPE and icLASER data to predict RNA polyA sites transcriptome-wide. a. Schematic of polyadenylation sequencing (PASseq). **b.** Genome browser track showing read density near the 3'-end of an RNA (RBFOX2) and demonstrating PAS seq specificity. **c.** Cumulative read density at 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 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.

556

557 **References**.

558 1. S. L. Klemm, Z. Shipony, W. J. Greenleaf, Chromatin accessibility and the regulatory
559 epigenome. *Nat Rev Genet* 10.1038/s41576-018-0089-8 (2019).

560 2. V. I. Risca, W. J. Greenleaf, Unraveling the 3D genome: genomics tools for multiscale
561 exploration. *Trends Genet* **31**, 357-372 (2015).

562 3. M. Kubota, C. Tran, R. C. Spitale, Progress and challenges for chemical probing of RNA
563 structure inside living cells. *Nat Chem Biol* **11**, 933-941 (2015).

564 4. E. J. Strobel, A. M. Yu, J. B. Lucks, High-throughput determination of RNA structures.
565 *Nat. Rev. Genet.* **19**, 615-634 (2018).

566 5. S. A. Mortimer, M. A. Kidwell, J. A. Doudna, Insights into RNA structure and function
567 from genome-wide studies. *Nat. Rev. Genet.* **15**, 469-479 (2014).

568 6. P. Tijerina, S. Mohr, R. Russell, DMS footprinting of structured RNAs and RNA-protein
569 complexes. *Nat. Protoc.* **2**, 2608-2623 (2007).

570 7. D. Mitchell, 3rd *et al.*, Glyoxals as in vivo RNA structural probes of guanine base-pairing.
571 *Rna* **24**, 114-124 (2018).

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

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

576 10. R. C. Spitale *et al.*, Structural imprints in vivo decode RNA regulatory mechanisms.
577 *Nature* **519**, 486-+ (2015).

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

581 12. T. Adilakshmi, R. A. Lease, S. A. Woodson, Hydroxyl radical footprinting in vivo:
582 mapping macromolecular structures with synchrotron radiation. *Nucleic Acids Res.* **34**,
583 e64 (2006).
584 13. C. Feng *et al.*, Light-activated chemical probing of nucleobase solvent accessibility
585 inside cells. *Nat Chem Biol* **14**, 325 (2018).
586 14. V. Tarwade, O. Dmitrenko, R. D. Bach, J. M. Fox, The Curtius rearrangement of
587 cyclopropyl and cyclopropenoyl azides. A combined theoretical and experimental
588 mechanistic study. *J. Org. Chem.* **73**, 8189-8197 (2008).
589 15. M. Albuszis, P. J. Roth, W. Pauer, H. U. Moritz, Two in one: use of azide functionality for
590 controlled photo-crosslinking and click-modification of polymer microspheres. *Polymer*
591 *Chemistry* **7**, 5414-5425 (2016).
592 16. L. B. Xu, J. Farrell, R. G. Karunakaran, A. Honglawan, S. Yang, Synthesis of dual-
593 functional copolymer with orthogonally photosensitive groups. *Journal of Polymer*
594 *Science Part a-Polymer Chemistry* **51**, 1215-1222 (2013).
595 17. C. Feng *et al.*, Light-activated chemical probing of nucleobase solvent accessibility
596 inside cells. *Nat Chem Biol* **14**, 276-283 (2018).
597 18. R. C. Spitale *et al.*, Structural imprints in vivo decode RNA regulatory mechanisms (vol
598 519, pg 486, 2015). *Nature* **527**, 264-264 (2015).
599 19. M. M. Abdelsayed *et al.*, Multiplex Aptamer Discovery through Apta-Seq and Its
600 Application to ATP Aptamers Derived from Human-Genomic SELEX. *ACS Chem Biol*
601 **12**, 2149-2156 (2017).
602 20. Y. Ding *et al.*, In vivo genome-wide profiling of RNA secondary structure reveals novel
603 regulatory features. *Nature* **505**, 696-700 (2014).
604 21. R. C. Spitale *et al.*, Structural imprints in vivo decode RNA regulatory mechanisms.
605 *Nature* **519**, 486-490 (2015).

606 22. J. Boucas, Integration of ENCODE RNAseq and eCLIP Data Sets. *Methods Mol Biol*
607 **1720**, 111-129 (2018).
608 23. E. L. Van Nostrand *et al.*, Robust transcriptome-wide discovery of RNA-binding protein
609 binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508-514 (2016).
610 24. D. D'Amico *et al.*, The RNA-Binding Protein PUM2 Impairs Mitochondrial Dynamics and
611 Mitophagy During Aging. *Mol Cell* **73**, 775-787 e710 (2019).
612 25. N. J. Proudfoot, Ending the message: poly(A) signals then and now. *Genes Dev.* **25**,
613 1770-1782 (2011).
614 26. P. J. Shepard *et al.*, Complex and dynamic landscape of RNA polyadenylation revealed
615 by PAS-Seq. *Rna-a Publication of the Rna Society* **17**, 761-772 (2011).
616 27. J. Harrow *et al.*, GENCODE: The reference human genome annotation for The
617 ENCODE Project. *Genome Res.* **22**, 1760-1774 (2012).
618 28. Y. D. Sun *et al.*, Molecular basis for the recognition of the human AAUAAA
619 polyadenylation signal. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E1419-E1428 (2018).
620 29. J. Brown, M. Pirrung, L. A. McCue, FQC Dashboard: integrates FastQC results into a
621 web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* **33**,
622 3137-3139 (2017).
623 30. D. R. Zerbino *et al.*, Ensembl 2018. *Nucleic Acids Res* **46**, D754-D761 (2018).
624 31. C. A. Sloan *et al.*, ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**, D726-
625 732 (2016).
626