



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Robust statistical modeling improves sensitivity of high-throughput RNA structure probing experiments

Citation for published version:

Selega, A, Sirocchi, C, Iosub, I, Granneman, S & Sanguinetti, G 2017, 'Robust statistical modeling improves sensitivity of high-throughput RNA structure probing experiments' Nature Methods, vol. 14, no. 1, pp. 83-89. DOI: 10.1038/nmeth.4068

Digital Object Identifier (DOI):

[10.1038/nmeth.4068](https://doi.org/10.1038/nmeth.4068)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

Nature Methods

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



1 Robust statistical modeling improves sensitivity of high- 2 throughput RNA structure probing experiments

3 Alina Selega¹, Christel Sirocchi², Ira Iosub², Sander Granneman², Guido Sanguinetti^{1,2}

4 ¹University of Edinburgh, School of Informatics, EH8 9AB, Edinburgh, UK

5 ²University of Edinburgh, Centre for Synthetic and Systems Biology, EH9 3BF, Edinburgh, UK.

6 **Structure probing coupled with high-throughput sequencing holds the potential to revolu-**
7 **tionize our understanding of the role of RNA structure in regulation of gene expression. De-**
8 **spite major technological advances, intrinsic noise and high coverage requirements greatly**
9 **limit the applicability of these techniques. Here we describe a probabilistic modeling pipeline**
10 **which accounts for biological variability and biases in the data, yielding statistically in-**
11 **terpretable scores for the probability of nucleotide modification transcriptome-wide. We**
12 **demonstrate on two yeast data sets that our method has greatly increased sensitivity, en-**
13 **abling the identification of modified regions on many more transcripts compared with ex-**
14 **isting pipelines. It also provides confident predictions at much lower coverage levels than**
15 **previously reported. Our results show that statistical modeling greatly extends the scope and**
16 **potential of transcriptome-wide structure probing experiments.**

17 RNA structure plays a key role in regulating RNA stability, transcription, and mRNA transla-
18 tion rates. In order to identify novel RNA structural regulatory elements, chemical and enzymatic
19 structure probing is routinely used to interrogate RNA structure both *in vivo* and *in vitro*¹. Current
20 *in silico* RNA structure prediction programs rely on thermodynamic estimates to generate the most

21 likely secondary structure models. By incorporating data from structure probing experiments, the
22 accuracy of secondary and tertiary RNA structure prediction can be significantly improved^{2,3}.

23 Most chemical RNA structure probing methods rely on the formation of adducts or cleav-
24 age of the RNA backbone, using as probes dimethylsulfate (DMS) and SHAPE reagents such as
25 1M7 (1-methyl-7-nitroisatoic anhydride) and NAI (2-methylnicotinic acid imidazolide)^{4,5}. In all
26 of these methods, the reagents terminate reverse transcription (RT), enabling detection of the sites
27 of cleavage or modification by primer extension analyses, followed by mapping the RT drop-off
28 position back to the reference sequence. These methods can be combined with next-generation se-
29 quencing (NGS) to simultaneously probe thousands of RNA molecules, as well as very long RNAs,
30 in a single RT reaction. Insights obtained by these techniques include the largely unstructured state
31 of stress-responsive transcripts in yeast and plants^{6,7}. Recently, we developed the ChemModSeq
32 structure probing pipeline to gain deeper understanding of RNA structural changes in long riboso-
33 mal RNA precursors during ribosome assembly⁸.

34 NGS is certainly revolutionizing the RNA structure probing field, however, several data an-
35 alytic questions need to be addressed. Firstly, NGS is often plagued by sequencing representation
36 and coverage biases introduced during library preparation⁹. Identifying and correcting such biases
37 is essential to avoid erroneous interpretations, however, to our knowledge, current methods do not
38 address these issues. Secondly, statistical assessments must be informed by an analysis of inter-
39 replicate variability in both control and treatment samples. Except for Mod-seq¹⁰, current methods
40 do not exploit replicate information, and, as a result, their output scores are not readily statistically

41 interpretable, often requiring setting arbitrary thresholds and other post-processing. Finally, a ma-
42 jor question in the field concerns the coverage per nucleotide necessary to get reliable chemical
43 reactivity values. Partly as a result of unresolved statistical issues in handling variability, current
44 recommendations indicate that very high coverage levels are required^{10,11}, which are normally only
45 met for a handful of transcripts in transcriptome-wide experiments.

46 To tackle these important issues, we developed *BUM-HMM* (Beta-Uniform Mixture Hidden
47 Markov Model), a statistical machine learning pipeline for modeling NGS RNA structure probing
48 data. BUM-HMM uses inter-replicate variability to identify transcript regions that are significantly
49 more modified, incorporating coverage and sequence bias information within the model. The out-
50 put of BUM-HMM is probabilistic, giving a transparent statistical interpretation which obviates the
51 need for arbitrary thresholds and post-processing. We demonstrate that BUM-HMM is highly sen-
52 sitive and remarkably robust even at low coverage, greatly improving over existing bioinformatic
53 pipelines.

54 **Results**

55 To demonstrate the strength of the BUM-HMM method, we re-analyzed high-throughput DMS and
56 1M7 RNA structure probing experiments performed on yeast 40S ribosomes⁸. This study gener-
57 ated biological triplicates of each chemical probing experiment with very high sequence coverage,
58 both in treatment and control samples (Supplementary Table 1). As secondary structure models for
59 rRNAs and crystal structures of yeast ribosomes are now readily available^{12,13}, these data allowed

60 us to investigate the sensitivity and specificity of BUM-HMM compared to existing methods. In
61 addition, we also generated two *in vivo* yeast mRNA transcriptome data sets using NAI as chemical
62 probe (see Methods for details), which enabled us to test the performance of BUM-HMM in the im-
63 portant context of a transcriptome-wide mRNA structure probing experiment. For these analyses,
64 between 36 and 55 million paired cDNA sequences were analyzed per sample (see Supplementary
65 Table 1 and Methods for details).

66 **Data preparation and model**

67 All cDNA libraries were generated by random priming^{6,8,11,14} and paired-end sequenced (see
68 Methods and Supplementary Fig. 1 for details). Paired-end sequencing allows normalization for
69 different read depths through calculating drop-off rates, which we define as the total number of
70 reads stopping at a nucleotide divided by the total number of reads that cover that nucleotide^{8,14}.
71 The full procedure is described in detail in Methods and schematically illustrated in Fig. 1.

72 Briefly, we quantify biological variability using the log-ratio between the drop-off rates at
73 the same nucleotide in a pair of control replicates (*log dor ratio*, LDR), for all possible pairs.
74 We assemble all control LDRs in a null distribution (Step A) and correct sequence and coverage
75 biases (Step B) to control for confounders (see Methods and Supplementary Fig. 2 for details). We
76 then evaluate empirical *p*-values for all treatment-control LDRs at each nucleotide (Step C) and
77 model these *p*-values using a Beta-Uniform mixture hidden Markov model (Step D) with hidden
78 states corresponding to presence or absence of modification (see Methods and Supplementary Fig.
79 3 for a theoretical justification of the Beta-Uniform choice). We use BUM-HMM to compute

80 posterior probabilities of chemical modification for all nucleotides (Step E), providing a robust
81 and statistically interpretable readout.

82 It is important to remark at this stage that, while single molecules are either modified or not at
83 a particular locus, interpreting structure probing data as binary may appear overly simplistic. Tran-
84 scripts *in vivo* exhibit dynamic secondary structures and may be bound by different proteins, so
85 that different molecules of the same transcript may be accessible to chemical reagents at different
86 positions. Furthermore, not all accessible nucleotides will be modified at low reagent concentra-
87 tions, such as those typically used in structure probing experiments. The correct interpretation
88 of the probabilistic output of BUM-HMM is therefore not that all transcript molecules with high
89 posterior probability at a locus are in a specific state of accessibility, but that the proportion of
90 modified molecules is sufficiently large to lead to an LDR value which cannot be explained by
91 random variability alone.

92 **Performance comparisons**

93 Interpreting and evaluating the outcome of structure probing experiments is a notoriously difficult
94 task due to a lack of “ground truth” examples to validate model predictions (see also Discus-
95 sion). In this respect, yeast 18S ribosomal RNA represents an important case of a high abundance
96 transcript with a well-defined and very stable secondary structure. We therefore first evaluated
97 BUM-HMM’s performance in terms of recovering the 18S structure from a recently published
98 chemical probing data set⁸. These data sets have extremely high coverage (with a mean coverage
99 per nucleotide close to 1 million for some samples, Supplementary Table 1), which clearly cannot

100 be achieved on many transcripts in transcriptome-wide studies. We thus later examine the per-
101 formance of BUM-HMM on the transcriptome data set, which reflects a more realistic coverage
102 scenario. We demonstrate through a number of case studies how BUM-HMM can aid the use of
103 structure prediction algorithms and recover structural features in conserved areas of transcripts, as
104 well as examine the robustness of BUM-HMM towards variations in coverage.

105 **BUM-HMM demonstrates state-of-the-art performance recovering the structure of 18S with**
106 **readily interpretable output**

107 Guided by the available 80S and 40S structures^{12,13}, we determined which nucleotides were ac-
108 cessible and single-stranded and should, in theory, be therefore modified by 1M7 or DMS. Notice
109 that this crystallographic structure is different from the phylogenetic (predicted) structure used in
110 other studies¹⁵. As DMS preferentially reacts with A's and C's, we were able to examine the sen-
111 sitivity and specificity of BUM-HMM. From many existing bioinformatic approaches^{6-8,10,14,16},
112 we chose the following methods to compare our model to: structure-seq⁶, Δ TCR¹⁴, which was
113 the strongest performer in a recent review¹⁶, and Mod-seq¹⁰, which to our knowledge is the only
114 method supporting multiple biological replicates. We evaluated all methods using the *receiver*
115 *operating characteristic* (ROC), which plots the false positive rate against the true positive rate
116 for different discrimination thresholds. A random predictor would have the *area under the ROC*
117 *curve* (the AUC statistic) equal to 0.5 and a higher value of the AUC indicates better performance.
118 When evaluated against the known crystal structure, BUM-HMM and Δ TCR were clearly the best

119 performers with AUC of 0.73 and 0.74, outperforming structure-seq and Mod-seq scoring at 0.68
120 and 0.64, respectively. The 1M7 data set demonstrated similar performance between methods
121 (Supplementary Table 2).

122 However, the dynamic output ranges of the methods vary dramatically; to enable compar-
123 isons with BUM-HMM while taking into account these differences, we separately examined the
124 true positive and true negative rate for different discrimination thresholds (scaling the scores to
125 range between 0 and 1). BUM-HMM demonstrated a 20% increase of the true positive rate
126 throughout most of the dynamic range compared to the other methods, for only a small decrease
127 of the true negative rate (Fig. 2a and 2b).

128 Fig. 2c shows the proportions of nucleobases called as modified by all methods, when dis-
129 criminating the scores at low, medium, and high thresholds or considering all scores greater than
130 zero. BUM-HMM has excellent specificity to A's and C's throughout its dynamic range. On the
131 contrary, structure-seq and Δ TCR do not discriminate as well between C's, G's, and U's when
132 considering all scores, demonstrating their reliance on arbitrary thresholds as the means to remove
133 noise. BUM-HMM identifies over a hundred modified nucleotides with very high posterior prob-
134 abilities, many more than the other methods do when considering high reactivity thresholds. It is
135 interesting to observe that on the 18S DMS data, BUM-HMM generates an almost binary output,
136 with few values between 0 and 1. This reflects the stability of the 18S transcript clearly evident
137 from the data, rather than a property of the model: BUM-HMM generates many more intermediate
138 values on the transcriptome data set.

139 Fig. 2d shows a fragment of the 18S secondary structure as predicted by BUM-HMM, with
140 many single-stranded A's and C's correctly identified. The results for all methods are shown on the
141 18S secondary structure models in Supplementary Fig. 4.

142 **BUM-HMM output aids computational prediction of secondary structures**

143 As explained earlier, the output posterior probabilities of BUM-HMM should not be directly
144 interpreted as secondary structure readouts in general. These probabilities can, however, pro-
145 vide valuable constraints to energy-based structure prediction software, such as RNAstructure¹⁷,
146 ViennaRNA¹⁸, and others. Such software predicts secondary structures of transcripts by minimiz-
147 ing the free energy associated with a particular “sequence–structure” configuration. For all but the
148 shortest transcripts, this is a very difficult combinatorial optimization problem, resulting in many
149 nearly equivalent optima corresponding to different structures. Transcripts *in vivo* are highly dy-
150 namic and can therefore exist in many different such configurations. However, under physiological
151 constraints, it can be expected that only a subset of all possible structures (from a free energy point
152 of view) will be present. We therefore used the BUM-HMM output as constraints for structure
153 prediction with the RNAstructure Web Server¹⁷.

154 To quantify the improvement provided by the BUM-HMM constraints, we selected as rep-
155 resentative examples the SCM4, RPL37A, and RPL19B genes (coding sequence only), which
156 encode a mitochondrial outer membrane protein and ribosomal 60S subunit proteins, correspond-
157 ingly. These genes all have good coverage levels (mean coverage per nucleotide 799, 38711, and

158 15798), thus avoiding problems with missing information; they are also relatively long transcripts
159 (564, 260, and 568 nucleotides long), and hence challenging for structure prediction algorithms.
160 We used the *Fold*¹⁷ method in RNAstructure to predict their secondary structure, with and without
161 the BUM-HMM constraints. *Fold* returns an ensemble of generally around 20 low free-energy
162 structures and we quantify the distance between two structures by using the binary Hamming dis-
163 tance. Constraining the algorithm with the BUM-HMM output considerably narrowed down the
164 search space for free-energy minimization, as demonstrated by smaller Hamming distances be-
165 tween the resulting structures (Fig. 3a, 3b, and 3c). Further, these structures were more similar to
166 the output of the alternative method *MaxExpect*¹⁷ compared to only using sequence (Supplemen-
167 tary Fig. 5). We conclude that using posterior probabilities generated by BUM-HMM as algorithm
168 constraints can improve secondary structure prediction for relatively long transcripts.

169

170 **BUM-HMM correctly predicts structure of conserved regions in U3 snoRNA**

171 While transcripts may co-exist in several different structural configurations, it is likely that some
172 of their sections present increased structural stability for correct cellular functioning (e.g. in order
173 to be bound by proteins). It is reasonable to expect highly conserved regions of a transcript to rep-
174 resent its more stable parts. To validate our model in the scenario of a more realistic transcriptome-
175 wide coverage, we turned to the small nucleolar RNA U3. U3 is a model gene for evolutionary
176 fitness studies¹⁹ and has an accepted secondary structure in yeast²⁰, making it a good candidate for
177 validation.

178 Even though the coverage on U3 was uneven and did not allow structural predictions on
179 the whole molecule, BUM-HMM achieved the AUC of 0.76 when evaluated on the highly con-
180 served regions located in boxes A, A', B, C, C', and D. Furthermore, when considering the longest
181 conserved region with 16 nucleotides (box A and one highly conserved upstream nucleotide),
182 BUM-HMM demonstrated excellent prediction accuracy of 0.88.

183 **BUM-HMM has increased informativeness on transcriptome-wide analysis of RNA structure**
184 **probing data**

185 To evaluate the applicability of the methods in the transcriptome-wide scenario, we generated
186 synthetic data sets by randomly selecting subsets of reads from the 18S DMS data set and evaluated
187 the consistency of the methods at lower coverage (see Methods for full details). BUM-HMM
188 showed excellent consistency as the mean coverage along the transcript was progressively reduced
189 (Fig. 4a), retaining significantly above random accuracy even at a reduction of almost 2000 times
190 (Supplementary Fig. 6). This performance challenges recent recommendations for the minimum
191 coverage level for chemical probing experiments¹¹, indicating that BUM-HMM can obtain reliable
192 predictions on a large fraction of transcripts in a standard transcriptomic experiment. Mod-seq and
193 structure-seq exhibited considerably lower levels of consistency (Fig. 4c and 4d) and behaved as
194 random predictors at the lowest coverage level. Highly consistent reactivity scores generated by
195 Δ TCR (Fig. 4b) were largely due to its extreme conservatism at the chosen threshold of 50% of
196 the dynamic range, at which it called no more than 20 nucleotides at all coverage levels. Notably,

197 all methods identified fewer modified nucleotides than BUM-HMM both on the full data set and
198 at all coverage levels, this difference being particularly striking with Δ TCR and Mod-seq (Fig. 4b
199 and 4c).

200 While performance analysis is hampered by a lack of a “ground truth” for most transcripts, a
201 more general assessment of the informativeness of the methods’ outputs is possible and instructive.
202 We therefore quantified how many transcripts had at least 5% of their length called as modified by
203 BUM-HMM and Δ TCR. We considered to be “called as modified” those nucleotides which ob-
204 tained a score above 50% of the dynamic range of the model (having removed outliers for Δ TCR).
205 With this procedure, BUM-HMM identified 2219 transcripts, while Δ TCR only retrieved 285. The
206 low number of transcripts identified by Δ TCR is at odds with previous studies^{6,7} suggesting that
207 many RNAs are largely accessible and unstructured *in vivo*; this conservativeness may be due to
208 the normalization procedures of Δ TCR¹⁴ (see Supplementary Fig. 7 for illustration of associated
209 problems).

210 We next analyzed the distribution of posterior probabilities across those mRNA transcripts
211 which had a non-zero score attached to more than 75% of their length, which we call effectively
212 probed. BUM-HMM selected 363 mRNA genes (Fig. 5a), which is in striking contrast with
213 Δ TCR’s 43 selected transcripts. When relaxing this criterion to (still highly informative) effective
214 probing of more than 50% of the length, the number of mRNAs selected by BUM-HMM increased
215 dramatically to 1764. Analyses of the 363 selected genes revealed that many appeared to have
216 long segments of almost completely unstructured regions (such as TDH3, Fig. 5b) and many had

217 significant structure in the coding sequence (such as YOR365W, Fig. 5b). We next calculated
218 the average FPKMs for these genes using the read counts from the control and treated sequencing
219 data. This revealed a broad distribution with a median 191 (Fig. 5b) and the lowest FPKM of
220 60 (YOR385W, Fig. 5b and 5c). This gene had an average coverage of 335 reads per nucleotide,
221 which we propose can be an indicative guideline of the lower bound on coverage required for
222 high-throughput RNA structure probing experiments to effectively probe long transcripts.

223 **Metabolic transcripts are generally flexible around the translation start site**

224 Structure in untranslated regions (UTR) and around the translation start site (AUG) can reduce
225 translation efficiency^{21,22}. Recent high-throughput RNA structure probing also revealed a weak
226 but significant negative correlation between RNA structure at that AUG *in vitro* and ribosome
227 occupancy²³. To test whether RNA structure measured *in vivo* also correlates with ribosome oc-
228 cupancy, we plotted the distribution of posterior probabilities around the translation start sites and
229 performed a *k*-means clustering to identify patterns in the data. This revealed five clusters with dif-
230 ferent reactivity profiles (Fig. 5d). For the majority of transcripts, the region around the AUG had
231 high posterior probabilities and therefore appeared to be largely unstructured (genes in clusters 0,
232 2, 3, 4). Interestingly, KEGG pathway analyses revealed that these clusters were highly enriched
233 for transcripts encoding for ribosomal and metabolic proteins, in particular proteins involved in
234 glycolysis/gluconeogenesis and amino acid biosynthesis (Supplementary Table 3). Remarkably,
235 the more structured transcripts in cluster 1 were mostly enriched for transcripts encoding proteins

236 involved in mitochondrial translation (Supplementary Table 3).

237 One possible explanation for why the metabolic transcripts appear largely unstructured *in*
238 *vivo* could be because they were occupied by ribosomes, which have an intrinsic RNA helicase
239 activity to unfold structured regions within mRNAs²⁴. We therefore asked whether there was a
240 significant correlation between RNA flexibility within that region and ribosome occupancy on
241 the transcripts. To test this, we calculated \log_2 of the sum of posterior probabilities within 50
242 nucleotides around the AUG and compared it to the translational efficiency obtained from the
243 recently published polysome microarray data²⁵ (Fig. 5e). This revealed that flexibility around the
244 AUG did not positively correlate with polysome occupancy (Pearson correlation: -0.196, *p*-value
245 = 0.0014). Similar results were obtained when using the entire 5' UTR region (Fig. 5f). Taken
246 together, these results suggest that high ribosome occupancy alone is not sufficient to explain why
247 certain transcripts were highly flexible in our *in vivo* NAI chemical probing data.

248 Discussion

249 High-throughput probing of RNA secondary structure offers unprecedented opportunities to eluci-
250 date the role of RNA structure in many fundamental biological processes. While the experimental
251 platforms are rapidly reaching maturity, several data analytic issues hinder their wider applicability
252 and adoption.

253 Our statistical pipeline addresses a number of such important problems. Firstly, it explic-
254 itly models the biological variability of the data, providing a statistical basis for determining the

255 significance of the observed signal. As such, it removes the need to set arbitrary thresholds and
256 perform extensive post-processing of the analysis results, yielding a clean and statistically inter-
257 pretable pipeline. This is in contrast to most existing methods and is a direct consequence of the
258 probabilistic formulation of BUM-HMM. In this respect, it is indebted to earlier probabilistic mod-
259 els of SHAPE-Seq data²⁶; notably, however, recent developments in the experimental technology,
260 and in particular, the shift to random-primed experimental designs, force a major change in model
261 architecture and motivate the non-parametric approach we take.

262 Our analysis identified important biases in the technology, especially prominent transcriptome-
263 wide, which can have severe downstream consequences in any analysis. While random-priming
264 designs effectively resolve the 3' biases of earlier SHAPE technologies, significant sequence and
265 coverage biases remain. Our method provides automated empirical strategies for correcting these
266 biases, potentially greatly extending the applicability of the technology.

267 Finally, the BUM-HMM model generates accurate and more informative results compared to
268 other methods. Crucially, its predictions remain consistent with reduced coverage, demonstrating
269 that the choice of an appropriate modeling framework can greatly increase the robustness of the
270 technology. This is borne out by the effectiveness of BUM-HMM on a transcriptome data set with
271 relatively low coverage: while current state-of-the-art methods can only provide information over
272 a handful of transcripts, BUM-HMM selected more than 360 transcripts, some of which had a
273 per nucleotide coverage as low as 335, heralding the advent of truly transcriptome-wide structure
274 probing experiments.

275 While BUM-HMM addresses many of the data analytic challenges associated with structure
276 probing data, it is important to stress that significant issues remain unsolved with the interpretation
277 of such data. Many factors may affect accessibility (protein binding being a prime example), and
278 in general transcripts *in vivo* may co-exist in multiple configurations, cautioning against simplistic
279 interpretations in terms of secondary structure. How structure probing data may be used to in-
280 form model-based structure prediction is an important and active research field^{27,28}. Our results
281 show that BUM-HMM constraints, when incorporated in structure prediction algorithms, lead to
282 more consistent structure models for many transcripts, demonstrating the importance of statisti-
283 cally sound data analytic strategies for downstream analyses.

284 **Accession codes.** The 18S rRNA (DMS and 1M7) and transcriptome-wide chemical probing sequencing
285 data are available in the Gene Expression Omnibus under accession numbers GSE52878 and GSE78208,
286 respectively.

287 **Code availability.** All of the code used in this study can be accessed in the following BitBucket reposi-
288 tory: https://bitbucket.org/aselega/bum_hmm_pipeline. The BUM-HMM pipeline will
289 be made available as a Bioconductor software package in due course.

290 **Acknowledgements** We thank all the members of the Granneman and Sanguinetti labs for critically read-
291 ing the manuscript. This work was supported by grants from the Wellcome Trust to S.G. (091549) and
292 I.I. (102334), a European Research Council grant to G.S. (MLC306999) and the Wellcome Trust Centre
293 for Cell Biology core grant (092076). A.S. is supported by grants from the UK Engineering and Physical
294 Sciences Research Council, Biological Sciences Research Council, and the UK Medical Research Council

295 (EP/F500385/1 and BB/F529254/1). Next generation sequencing was carried out by Edinburgh Genomics,
296 The University of Edinburgh. Edinburgh Genomics is partly supported through core grants from NERC
297 (R8/H10/56), MRC (MR/K001744/1) and BBSRC (BB/J004243/1).

298 **Contributions** All authors contributed to planning the experiments and computational procedures. C.S.,
299 I.I., and S.G. carried out the experiments. G.S. and A.S. developed the computational analysis pipeline.
300 A.S., C.S., S.G., and G.S. performed the bioinformatics and computational analyses of the sequencing data.
301 All authors contributed to writing the manuscript and approved the final manuscript.

302 **Competing Interests** The authors declare that they have no competing financial interests.

303 **Correspondence** Correspondence and requests for materials should be addressed to:

304 S.G. sgrannem@staffmail.ed.ac.uk.

305 G.S. gsanguin@inf.ed.ac.uk.

306 **References**

- 307 1. Kubota, M., Tran, C. & Spitale, R. C. Progress and challenges for chemical probing of RNA
308 structure inside living cells. *Nature chemical biology* **11**, 933–941 (2015).
- 309 2. Wu, Y. *et al.* Improved prediction of RNA secondary structure by integrating the free energy
310 model with restraints derived from experimental probing data. *Nucleic acids research* **43**,
311 7247–7259 (2015).

- 312 3. Ouyang, Z., Snyder, M. P. & Chang, H. Y. SeqFold: genome-scale reconstruction of RNA
313 secondary structure integrating high-throughput sequencing data. *Genome research* **23**, 377–
314 387 (2013).
- 315 4. Mortimer, S. A. & Weeks, K. M. A fast-acting reagent for accurate analysis of RNA secondary
316 and tertiary structure by SHAPE chemistry. *Journal of the American Chemical Society* **129**,
317 4144–4145 (2007).
- 318 5. Spitale, R. C. *et al.* RNA SHAPE analysis in living cells. *Nature chemical biology* **9**, 18–20
319 (2013).
- 320 6. Ding, Y. *et al.* In vivo genome-wide profiling of rna secondary structure reveals novel regula-
321 tory features. *Nature* **505**, 696–700 (2014).
- 322 7. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing
323 of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705
324 (2013).
- 325 8. Hector, R. D. *et al.* Snapshots of pre-rRNA structural flexibility reveal eukaryotic 40S assem-
326 bly dynamics at nucleotide resolution. *Nucleic acids research* gku815 (2014).
- 327 9. van Dijk, E. L., Jaszczynszyn, Y. & Thermes, C. Library preparation methods for next-
328 generation sequencing: tone down the bias. *Experimental cell research* **322**, 12–20 (2014).
- 329 10. Talkish, J., May, G., Lin, Y., Woolford, J. L. & McManus, C. J. Mod-seq: high-throughput
330 sequencing for chemical probing of rna structure. *rna* **20**, 713–720 (2014).

- 331 11. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery
332 by SHAPE and mutational profiling (SHAPE-MaP). *Nature methods* (2014).
- 333 12. Ben-Shem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science*
334 (*New York, N.Y.*) **334**, 1524–1529 (2011).
- 335 13. Aylett, C. H. S., Boehringer, D., Erzberger, J. P., Schaefer, T. & Ban, N. Structure of a yeast
336 40S-eIF1-eIF1A-eIF3-eIF3j initiation complex. *Nature structural & molecular biology* **22**,
337 269–271 (2015).
- 338 14. Kielpinski, L. J. & Vinther, J. Massive parallel-sequencing-based hydroxyl radical probing of
339 RNA accessibility. *Nucleic acids research* **42**, e70 (2014).
- 340 15. Tang, Y. *et al.* Structurefold: genome-wide rna secondary structure mapping and reconstruc-
341 tion in vivo. *Bioinformatics* **btv213** (2015).
- 342 16. Kielpinski, L. J., Sidiropoulos, N. & Vinther, J. Reproducible Analysis of Sequencing-Based
343 RNA Structure Probing Data with User-Friendly Tools. *Methods in enzymology* **558**, 153–180
344 (2015).
- 345 17. Reuter, J. S. & Mathews, D. H. Rnastructure: software for rna secondary structure prediction
346 and analysis. *BMC bioinformatics* **11**, 1 (2010).
- 347 18. Lorenz, R. *et al.* Viennarna package 2.0. *Algorithms for Molecular Biology* **6**, 1 (2011).
- 348 19. Puchta, O. *et al.* Network of epistatic interactions within a yeast snorna. *Science* **352**, 840–844
349 (2016).

- 350 20. Méreau, A. *et al.* An in vivo and in vitro structure-function analysis of the *Saccharomyces*
351 *cerevisiae* u3a snorncp: protein-rna contacts and base-pair interaction with the pre-ribosomal
352 rna. *Journal of molecular biology* **273**, 552–571 (1997).
- 353 21. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of
354 gene expression in *Escherichia coli*. *Science (New York, N.Y.)* **324**, 255–258 (2009).
- 355 22. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by
356 both codon bias and folding energy. *Proceedings of the National Academy of Sciences of the*
357 *United States of America* **107**, 3645–3650 (2010).
- 358 23. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature*
359 **467**, 103–107 (2010).
- 360 24. Takyar, S., Hickerson, R. P. & Noller, H. F. mRNA Helicase Activity of the Ribosome. *Cell*
361 **120**, 49–58 (2005).
- 362 25. Arribere, J. A., Doudna, J. A. & Gilbert, W. V. Reconsidering movement of eukaryotic mRNAs
363 between polysomes and P bodies. *Molecular cell* **44**, 745–758 (2011).
- 364 26. Aviran, S. *et al.* Modeling and automation of sequencing-based characterization of rna struc-
365 ture. *Proceedings of the National Academy of Sciences* **108**, 11069–11074 (2011).
- 366 27. Deng, F., Ledda, M., Vaziri, S. & Aviran, S. Data-directed rna secondary structure prediction
367 using probabilistic modeling. *RNA* (2016).

368 28. Eddy, S. R. Computational analysis of conserved rna secondary structure in transcriptomes
369 and genomes. *Biophysics* **43** (2014).

370 **Figure 1: Overview of the BUM-HMM computational analysis pipeline.** (a) Null dis-
371 tribution of LDRs is computed for all pairs of control replicate samples, quantifying variability in
372 drop-off rate observed by chance. (b) Coverage-dependent bias is corrected by applying a vari-
373 ance stabilization transformation. For transcriptome-wide data sets, different null distributions are
374 computed for different nucleobase patterns to address sequence-dependent bias. (c) Per-nucleotide
375 empirical p -values are computed for all pairs of treatment and control replicate samples, by com-
376 paring the corresponding LDRs to the null distribution. (d) BUM-HMM is run on p -values as
377 observations, leaving out any nucleotides with missing data. (e) The output is a posterior probabili-
378 ty of modification, ranging from 0 to 1, for each nucleotide included in the analysis.

379 **Figure 2: BUM-HMM identifies many modified nucleotides of 18S ribosomal RNA with**
380 **high accuracy and specificity.** (a, b) True positive rate and true negative rate of all methods for
381 reconstructing secondary structure of 18S rRNA, evaluated against the known crystal structure. (c)
382 Base composition of called nucleotides for all methods, when considering scores greater than: a
383 value close to zero (10^{-6}), a low reactivity threshold (0.1), a medium reactivity threshold (0.4), and
384 a high reactivity threshold (0.85). (d) A fragment of the 18S secondary structure with bases colored
385 according to the BUM-HMM posterior probability at the corresponding nucleotide position.

386 **Figure 3: Using BUM-HMM output results in more consistent secondary structure pre-**

387 **diction.** (a) Distribution of Hamming distances between all pairs of secondary structures ($n = 20$)
388 predicted for SCM4 by *Fold* when using only sequence (blue) and adding the BUM-HMM output
389 as constraints (red), and a fragment of the lowest free energy structure. (b, c) Same as in (a), for
390 RPL37A (b) and RPL19B (c).

391 **Figure 4: BUM-HMM is highly consistent at low coverage and calls more nucleotides**
392 **modified.** (a) Consistency of posterior probabilities generated by BUM-HMM on data sets with
393 progressively lower mean coverage (shown on the x-axis), synthesized from the DMS data set for
394 18S rRNA (see Methods for details). For each coverage level, base composition of nucleotides
395 called as modified is shown in a corresponding barplot, averaged across 10 selections of subsets.
396 The barplot in a shaded rectangle corresponds to the base composition of called nucleotides on the
397 full data set. (b, c, d) Consistency of reactivity scores generated by Δ TCR (b), Mod-seq (c), and
398 structure-seq (d) on the same synthetic data sets, with prior outlier removal.

399 **Figure 5: Flexibility of 5' UTR and ribosome occupancy do not show a significant pos-**
400 **itive correlation *in vivo*.** (a) Distribution of posterior probabilities over 363 protein-coding tran-
401 scripts. The heatmap displays posterior probabilities for 363 mRNA transcripts that were selected
402 from the transcriptome-wide data. The mRNAs were sorted by length (from short to long) and ex-
403 tended at each end by 300 nucleotides. The two black lines indicate the position of the start codon
404 and stop codon, respectively. (b) Genome browser examples showing posterior probabilities of a
405 highly expressed gene (TDH3; average FPKM = 3491) and a lowly expressed gene (YOR385W;
406 average FPKM = 60). (c) Violin plot showing the distribution of average FPKMs, calculated using

407 the sequence reads from the control and NAI datasets. **(d)** Many transcripts are flexible around
408 the translation start site. The plot shows the distribution of posterior probabilities 50 nucleotides
409 around the translation start site (AUG). *K*-means clustering revealed five clusters with different
410 distributions of probabilities. On the right side of the heat map, cumulative plots are shown for
411 each cluster. The number of genes (*n*) in each cluster is also indicated. **(e)** High structural flexi-
412 bility does not correlate with high ribosome occupancy. For each gene, we calculated \log_2 of the
413 sum of posterior probabilities from the heat map data shown in **(d)** and plotted it against the \log_2
414 of the reported enrichment of the transcript in polysomes²⁵. **(f)** Same as in **(e)** but with the entire
415 5' UTR.

416 **Methods**

417 **ChemModSeq library preparation.**

418 The 18S DMS and 1M7 data sets were previously described⁸. To generate the NAI transcriptome-
419 wide data set, yeast cells (BY4741 strain) were grown to exponential phase and harvested by cen-
420 trifugation. Cells were subsequently resuspended in 1 volume of phosphate buffer saline (PBS).
421 NAI (dissolved in DMSO) was added to the suspension in a final concentration of 100 mM (5%
422 DMSO final) and incubated for 10 minutes at room temperature. Cells were harvested by cen-
423 trifugation, washed with ice-cold PBS and snap-frozen in liquid nitrogen. Total RNA was ex-
424 tracted as previously described²⁹. The mRNAs were isolated using the PolyATtract mRNA iso-
425 lation kit, according to manufacturer's procedures (Promega). Two biological replicates were
426 generated for the transcriptome-wide analyses. The ChemModSeq libraries were generated as

427 previously described⁸. Briefly, cDNA was generated by random priming using a random hexamer
428 oligo⁸. Subsequently, a DNA adapter was ligated to the 3' end of cDNAs using CircLigase. These
429 adapters contained a random nucleotide at the 5' end to minimize the sequence representation bi-
430 ases introduced during the linker ligation reaction. Following PCR, libraries were resolved on 2%
431 Metaphor gels and fragments between 200-700 were gel purified. Samples were sequenced on
432 Illumina HiSeq2500 systems.

433 **Sequence data processing and raw data analysis.**

434 To process the fastq files the pyCRAC package was used³⁰. To demultiplex the raw sequencing
435 data we used pyBarcodeFilter.py, after which the remaining random nucleotide was removed from
436 the 5' end of the forward reads. The data were subsequently collapsed using pyFastqDuplicateRe-
437 mover.py that utilizes the random barcode information present in the 5' adapters to remove poten-
438 tial PCR duplicates. The resulting fasta file was mapped to the *Saccharomyces cerevisiae* genome
439 (version R64, ENSEMBL) using novoalign 2.05 and only uniquely mapped reads were considered.
440 PyReadCounters.py was subsequently used to generate read counts and FPKMs for all annotated
441 features. The resulting GTF output files were converted to tab-delimited files containing three
442 columns: chromosome, genomic position, and coverage or drop-off counts using pyGTF2sgr.py.
443 These files were then fed to the BUM-HMM model to generate posterior probabilities.

444 **Data characterization.**

445 Using the final output files (see Sequence data processing and raw data analysis), the drop-off rate
446 was computed for all nucleotide positions in each replicate as a measure of nucleotide's reactivity
447 to the probing reagent in a given experiment. By definition, the drop-off rate ranges between 0 and

448 1. All drop-off rates were normalized to a common median across replicate samples.

$$\text{drop-off rate} = \frac{\text{drop-off count}}{\text{coverage}}$$

449 A measure of inter-replicate variability at each nucleotide position is defined as the log-ratio of
450 drop-off rates (LDR) in a pair of replicate samples i and j :

$$\log \left(\frac{\text{drop-off rate}_i}{\text{drop-off rate}_j} \right) = \log (\text{drop-off rate}_i) - \log (\text{drop-off rate}_j)$$

451 If the drop-off rates are similar in both samples, the LDR will be close to 0, indicating little vari-
452 ability. In contrast, different drop-off rates would result in an LDR large in absolute value. LDRs
453 in control conditions collectively describe the variability in drop-off rates that could be observed
454 in the absence of the probing reagent. The set of these define the *null distribution* of LDRs.

455 LDRs are then computed for each combination of treatment-control replicates, quantifying
456 the difference between the drop-off rate observed in a treatment experiment with respect to a con-
457 trol replicate. These are compared to the null distribution giving rise to empirical p -values. For
458 efficiency, LDRs are compared to the precomputed quantiles of the null distribution. The p -value of
459 an LDR represents the probability of it being insignificantly different from what could be observed
460 by chance.

$$p\text{-value} = 1 - q, \text{ where } q \text{ is the closest quantile}$$

461 **Preprocessing.**

462 In order to use the log transform, it is necessary to ensure that no nucleotides have zero drop-off
463 rates. Therefore, only those nucleotides with non-zero drop-off counts for a corresponding pair

464 of replicate samples are used. The pipeline also features a user-defined parameter describing the
465 minimum level of coverage that nucleotides should have to be included in the analysis.

466 **Model.**

467 Empirical p -values, computed for each nucleotide position and each treatment-control comparison
468 (of which there are nm for n treatment and m control experimental replicates) are passed onto a
469 hidden Markov model. The model has a hidden state h_t ($t = 1 \dots T$ for T nucleotides) representing
470 the true binary state of the t -th nucleotide (modified, 1 or unmodified, 0) and the observed variable
471 v_t , corresponding to the empirical p -value at that position. P -values corresponding to different
472 pairs of treatment-control replicates are assumed to be independent measurements. Notice that,
473 since p -values are used as features and not for decision making, no issues of multiple hypothesis
474 testing arise.

475 Transition probabilities are defined through empirically derived lengths of single- and double-
476 stranded stretches of nucleotides. The model assumes expected uninterrupted stretches of 20
477 double-stranded or constrained nucleotides and 5 single-stranded or flexible nucleotides.

478 Emission probabilities come from a Beta-Uniform mixture (BUM) model. This design ex-
479 ploits the result that p -values are uniformly distributed under the null hypothesis³¹. P -values cor-
480 responding to accessible nucleotides are modeled with a Beta distribution, which favors small
481 values, accommodating the fact that accessible nucleotides would have LDRs greater than most
482 values in the null distribution. The p -value distribution computed for the transcriptome-wide data
483 set strongly agrees with this model (Supplementary Fig. 3). The HMM is run separately on con-

484 tinuous stretches of nucleotides with a user-specified minimum coverage threshold and a non-zero
 485 drop-off rate in at least one treatment sample.

$$p(v_t|h_t = 0) \sim U(0, 1)$$

486

$$p(v_t|h_t = 1) \sim \text{Beta}(\alpha, \beta), \text{ with } \alpha = 1, \beta = 10$$

487 **Optimization of parameters.**

488 We provide a strategy to optimize parameters of the Beta distribution with respect to the data. This
 489 strategy uses the expectation-maximization (EM) algorithm³² and Newton's optimization method.

490 The iterative EM-algorithm starts with the initial values of $\alpha = 1$ and $\beta = 10$, with which the
 491 posterior probabilities are computed. It then computes new estimates for α and β using Newton's
 492 optimization method. Newton's method finds the maximum of the expected complete data log-
 493 likelihood, or more precisely, its relevant terms. The shape parameters α and β only appear in the
 494 emission term and within that, only in the component corresponding to the modified state of the
 495 latent variable h_t .

496 The expected complete data log-likelihood is given by the following expression (all expecta-
 497 tions are with respect to corresponding distributions):

$$\langle \log p(v_{1:T}, h_{1:T} | \alpha, \beta) \rangle = \langle \log p(h_1) \rangle + \left\langle \sum_{t=1}^T \sum_{n=1}^N \log p(v_t^n | h_t) \right\rangle + \left\langle \sum_{t=1}^{T-1} \log p(h_{t+1} | h_t) \right\rangle,$$

498 for $t = 1 \dots T$ nucleotides and $n = 1 \dots N$ number of treatment-control comparisons. The relevant
 499 term corresponds to emission probabilities (second term in the previous expression):

$$\left\langle \sum_{t=1}^T \sum_{n=1}^N \log p(v_t^n | h_t) \right\rangle = \sum_{t=1}^T \sum_{n=1}^N \log p(v_t^n | h_t = 0) p(h_t = 0 | v_{1:T}^n) +$$

500

$$+ \sum_{t=1}^T \sum_{n=1}^N \log p(v_t^n | h_t = 1) p(h_t = 1 | v_{1:T}^n)$$

501 Within that expression, the relevant term corresponds to the modified state of the hidden variable

502 (second term in the previous expression):

$$\begin{aligned} \sum_{t=1}^T \sum_{n=1}^N \log p(v_t^n | h_t = 1) p(h_t = 1 | v_{1:T}^n) &= \sum_{t=1}^T \sum_{n=1}^N \gamma_t \log \frac{(v_t^n)^{\alpha-1} (1 - v_t^n)^{\beta-1}}{B(\alpha, \beta)} = \\ &= F, \end{aligned}$$

503

504

where $\gamma_t = p(h_t = M | v_{1:T}^n)$ is the responsibility.

505 The first order derivatives of F are:

$$\frac{\delta F}{\delta \alpha} = \sum_{t=1}^T \sum_{n=1}^N \gamma_t \log v_t^n - \gamma_t (\psi_0(\alpha) - \psi_0(\alpha + \beta))$$

506

$$\frac{\delta F}{\delta \beta} = \sum_{t=1}^T \sum_{n=1}^N \gamma_t \log (1 - v_t^n) - \gamma_t (\psi_0(\beta) - \psi_0(\alpha + \beta))$$

507 The second order derivatives of F are:

$$\frac{\delta^2 F}{\delta \alpha^2} = \sum_{t=1}^T \gamma_t N(\psi_1(\alpha + \beta) - \psi_1(\alpha))$$

508

$$\frac{\delta^2 F}{\delta \alpha \delta \beta} = \sum_{t=1}^T \gamma_t N \psi_1(\alpha + \beta)$$

509

$$\frac{\delta^2 F}{\delta \beta^2} = \sum_{t=1}^T \gamma_t N(\psi_1(\alpha + \beta) - \psi_1(\beta)),$$

510 where ψ is the polygamma function. Log transform is applied at the beginning of the algorithm511 to ensure that the estimated α and β are positive. Posterior probabilities are recomputed with the512 new estimates of α and β and the process is repeated a maximum number of 10 times or until the

513 parameter values stop changing within the small predefined tolerance range.

514 **Bias correction.**

515 We used the transcriptome-wide data set to identify potential confounding factors which influence
516 the LDRs in the absence of a reagent. The aim is to transform all LDRs accordingly and eliminate
517 the revealed biases.

518 **Coverage bias.**

519 The coverage bias was identified by plotting the control LDRs as a function of the inter-replicate
520 mean coverage at the corresponding nucleotide position (Supplementary Fig. 2a and 2b).

521 This bias is corrected by learning the functional dependency between these variables and
522 transforming the data to reduce the variance of LDRs. We model drop-off count as a binomially
523 distributed variable, which thus has the following standard deviation:

$$\sigma[\text{drop off count}] = \sqrt{np(\text{drop off})(1 - p(\text{drop off}))}, \text{ for a nucleotide covered } n \text{ times.}$$

524 Consequently, LDR has a standard deviation of:

$$\sigma[\text{LDR}] \propto \frac{\sigma[\text{drop off count}]}{n} = \frac{\sqrt{p(1 - p)}}{\sqrt{n}}$$

525 Therefore, the functional relationship between log-ratios and coverage can be modeled as a $k\frac{1}{\sqrt{n}} + b$,
526 with some unknown parameters k and b , which are learned from the data using a non-linear least
527 squares technique. Then, all LDRs are rescaled by this model with fitted parameters. For efficient
528 runtime on transcriptome-wide data sets, the LDRs are split in bins of equal coverage ranges and
529 the 95th quantile of LDRs and mean coverage are computed for each bin. These are used for

530 parameter fitting. Supplementary Fig. 2c and 2d show that the transformed LDRs have reduced
531 dependency on coverage.

532 **Sequence bias.**

533 We compared the resulting LDR null distributions when separately considering nucleobase patterns
534 of three (AAA, AAT, AAG, ...). For each of the 64 combinations of nucleobases, the transcriptome
535 sequence was searched for all places of its occurrence. The LDRs of the middle nucleotide at these
536 occurrences defined the null distribution specific to this nucleobase combination. Supplementary
537 Fig. 2e and 2f demonstrate significant differences between these null distributions.

538 To correct for this sequence-dependent bias, we store the quantiles of each of the 64 different
539 null distributions and compute empirical p -values by keeping track of which nucleobase triplet
540 corresponds to the current nucleotide position and looking up values from the corresponding null
541 distribution.

542 Due to the short length of the 18S ribosomal RNA molecule, the sequence-bias correcting
543 step was omitted from the analysis when handling the corresponding data sets.

544 **Handling of missing data and outliers.**

545 The methods used in the evaluation^{6,10,14} not only generate scores with drastically differing dy-
546 namic ranges, but also assume different interpretations of the same score values. For instance,

547 Δ TCR makes no distinction between the equal drop-off rates in control and treatment conditions
548 and no coverage, assigning a score of 0 in both cases. Structure-seq marks missing data with a
549 dummy value, whereas Mod-seq clamps the scenarios of no coverage and no significant modifi-
550 cation to the same score of 0. Further, the outputs of these methods have clear outliers, with a
551 handful of values being much larger than the 99th quantile of the output distribution. Therefore,
552 simply choosing the midpoint of the dynamic range for binarizing the resulting classifications
553 would result in as few as a single true positive for some methods.

554 Thus, when performing evaluation, we set the missing data (for those methods that use it)
555 and the outliers (computed as the values greater than the 99.5th quantile of the output distribution)
556 to 0. Considering other strategies, such as removing outliers or only evaluating on the non-missing
557 data, resulted in grossly limited outputs generated by some methods for the simulated low cover-
558 age levels. Our choice, while circumventing these problems and enabling comparisons, follows
559 the commonly utilized assumption that the reactivity of zero does not carry significant structural
560 information.

561 Overall, these difficulties expose the problems associated with the discussed methods; namely,
562 the absence of a unified output scale (which therefore leads to arbitrary threshold setting), gross
563 outliers, and inability to represent missing data, which thus results in extreme conservatism of the
564 classification. BUM-HMM addresses this by having a clearly defined probability output range and
565 separating out the nucleotides about which no predictions can be made.

566 When computing true positive and negative rates, the output scores of all methods were

567 normalized to the range of BUM-HMM. AUCs and true positive and negative rates were computed
568 with the ROCR package³³. When characterizing the methods' sensitivities using the DMS data
569 set specific to A's and C's, the outputs of Δ TCR and Mod-seq were normalized with the 2-8%
570 normalization rule³⁴ to enable comparisons at the same (previously used) low, medium, and high
571 reactivity thresholds^{34,35}.

572 **Secondary structure prediction.**

573 When generating secondary structures informed by BUM-HMM, posterior probabilities were up-
574 loaded to the RNAstructure Web Server¹⁷ as a SHAPE constraints file with default parameter values
575 used. For RPL37A and RPL19B, the structure was predicted for the longest CDS region.

576 **Performance evaluation of BUM-HMM on the conserved regions of U3 snoRNA.**

577 Conservation scores associated with the human U3 snoRNA were taken from Rfam³⁶. Highly con-
578 served parts of the box regions, matching in sequence between the human³⁷ and yeast transcripts²⁰,
579 were selected, with three lowly conserved nucleotides allowed in the middle of the regions (a to-
580 tal of 40 nucleotides). Evaluation was performed on those nucleotides with an attached posterior
581 probability $p > 0$ (28 of those nucleotides).

582 **Lower coverage simulation analysis.**

583 To evaluate the output consistency of the methods at lower coverage levels, we generated synthetic
584 data sets by randomly selecting subsets of 2 million, 1 million, 100000, 30000, 20000, 10000, and
585 1000 reads from the 18S DMS data set. For each subset, 10 such selections were made. Files
586 with coverage and drop-off counts were generated for each selection and passed to BUM-HMM.
587 Consistency was evaluated with the AUC statistic between the output scores generated by each
588 method for a given synthetic subset selection and the whole data set. For all methods, outliers
589 were handled as described above and calling of modified nucleotides (used for the barplots of base
590 composition) was performed at the threshold of 50% of the dynamic range of each method, after
591 having dealt with the outliers.

592 **Methods References**

- 593 29. Tollervey, D. A yeast small nuclear RNA is required for normal processing of pre-ribosomal
594 RNA. *The EMBO journal* **6**, 4169–4175 (1987).
- 595 30. Webb, S., Hector, R. D., Kudla, G. & Granneman, S. PAR-CLIP data indicate that Nrd1-
596 Nab3-dependent transcription termination regulates expression of hundreds of protein coding
597 genes in yeast. *Genome biology* **15**, R8 (2014).
- 598 31. Murdoch, D. J., Tsai, Y.-L. & Adcock, J. P-values are random variables. *The American*
599 *Statistician* (2012).

- 600 32. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data
601 via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* 1–38
602 (1977).
- 603 33. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. Rocr: visualizing classifier performance
604 in r. *Bioinformatics* **21**, 7881 (2005). URL <http://rocr.bioinf.mpi-sb.mpg.de>.
- 605 34. Low, J. T. & Weeks, K. M. Shape-directed rna secondary structure prediction. *Methods* **52**,
606 150–158 (2010).
- 607 35. Lucks, J. B. *et al.* Multiplexed RNA structure characterization with selective 2'-hydroxyl acy-
608 lation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National*
609 *Academy of Sciences of the United States of America* **108**, 11063–11068 (2011).
- 610 36. Nawrocki, E. P. *et al.* Rfam 12.0: updates to the rna families database. *Nucleic acids research*
611 *gku1063* (2014).
- 612 37. Granneman, S. *et al.* Role of pre-rna base pairing and 80s complex formation in subnucleolar
613 localization of the u3 snorncp. *Molecular and cellular biology* **24**, 8600–8610 (2004).

614 **Supplementary Information**

615 **Supplementary Table 1:** Overview of paired cDNA reads analyzed from each data set. All raw
616 sequencing data have been collapsed before aligning to the reference sequences to remove potential
617 PCR duplicates. Only properly paired reads were considered for the analyses.

618 **Supplementary Table 2:** Accuracy of reconstructing secondary structure of 18S ribosomal
619 RNA from the 1M7 data set for all methods, measured with the AUC statistic against the known
620 crystal structure of the rRNA.

621 **Supplementary Table 3:** KEGG pathway analysis of the k -means clusters shown in Fig. 4d.
622 These analyses were performed on the string-db server (www.string-db.org).

623 **Supplementary Figure 1: ChemModSeq library preparation design.** Chemically probed
624 RNAs were reverse transcribed with an oligonucleotide containing a random hexamer and an Il-
625 lumina compatible sequence for PCR amplification. Subsequently adapters were ligated to the 3'
626 end of cDNAs that contained six random nucleotides and a six nucleotide barcode followed by
627 another random nucleotide. The latter was introduced to minimize sequence bias representation
628 introduced during the CircLigase ligation reaction. The six random nucleotides were used to elim-
629 inate potential PCR duplicates. Indexing barcodes were added to the 3' adapter sequence by PCR.
630 The in-read barcodes in the 5' end of the PCR product were processed using pyBarcodeFilter.py
631 and reads were collapsed using pyFastqDuplicateRemover.py from the pyCRAC package³⁰.

632 **Supplementary Figure 2: Coverage- and sequence-dependent biases were identified**
633 **in the transcriptome data set.** (a, b) Presence of a coverage-dependent bias, reflected by the
634 dependency between the average LDR and the mean coverage at each nucleotide position in a
635 pair of control replicate samples, for all such pairs. (c, d) Same dependency plotted as in (a,
636 b) after applying a bias-correcting strategy to the LDRs. (e, f) Presence of a sequence-dependent
637 bias, reflected by differing null distributions of LDRs, each computed only for nucleotide positions

638 corresponding to a given trinucleotide pattern.

639 **Supplementary Figure 3: Distribution of empirical p -values for the transcriptome data**
640 **set closely follows a Beta-Uniform distribution on both strands.** The histograms show the
641 distributions of empirical p -values associated with LDRs between all combinations of treatment
642 and control samples on the transcriptome data set.

643 **Supplementary Figure 4: BUM-HMM correctly identifies many flexible A's and C's as**
644 **modified nucleotides.** Secondary structures of 18S ribosomal RNA with bases colored according
645 to the reactivity score or posterior probability at the corresponding nucleotide position, generated
646 by BUM-HMM, Δ TCR¹⁴, Mod-seq¹⁰, and structure-seq⁶ analysis pipelines.

647 **Supplementary Figure 5: Using BUM-HMM output results in more consistent sec-**
648 **ondary structure prediction across different methods. (a)** Distribution of Hamming distances
649 between the structures predicted for SCM4 by *Fold*¹⁷ ($n = 20$) and by *MaxExpect*¹⁷ ($n = 3$ with
650 sequence, $n = 1$ with BUM-HMM) when using only sequence (blue) and adding the BUM-HMM
651 output as constraints (red). **(b, c)** Same as in **(a)**, for RPL37A **(b)** and RPL19B **(c)** (with *Fold*,
652 $n = 20$ structures were generated, with *MaxExpect*, $n = 1$ structure).

653 **Supplementary Figure 6: BUM-HMM retains good accuracy at 18S secondary struc-**
654 **ture reconstruction at lower coverage levels.** Agreement with the 18S crystal structure of the
655 posterior probabilities generated by BUM-HMM on data sets with progressively lower mean cov-
656 erage (shown on the x-axis), synthesized from the DMS data set for 18S ribosomal RNA. For each

657 coverage level, the subsets of reads were randomly selected from the full data set 10 times.

658 **Supplementary Figure 7: The Δ TCR algorithm produces very high numbers in regions**

659 **with low coverage.** Shown is a genome browser image of a gene (YHB1) with an FPKM of 190.

660 The red-dotted box shows a region near the 3' end of the gene where there is low coverage. The top

661 two panels show the Δ TCR¹⁴ output, with the second panel displaying the same data but scaled to

662 a maximum Δ TCR value of 0.025. The third panel shows the BUM-HMM posterior probabilities

663 for the same region. The last four panels show the cDNA coverage over the gene from the two

664 control RNA sequencing data and the two NAI treated sequencing data.