BIOCHEMISTRY
including biophysical chemistry & molecular biology

Rapid Report

pubs.acs.org/biochemistry

# Standardization of RNA Chemical Mapping Experiments

Wipapat Kladwang,[†] Thomas H. Mann,[†] Alex Becka,[†] Siqi Tian,[†] Hanjoo Kim,[‡] Sungroh Yoon,[‡] and Rhiju Das*[,†,§]

[†]Department of Biochemistry, Stanford University, Stanford, California 94305, United States

[‡]Department of Electrical and Computer Engineering, Seoul National University, Seoul 151-744, Korea

[§]Department of Physics, Stanford University, Stanford, California 94305, United States

Supporting Information

**ABSTRACT:** Chemical mapping experiments offer powerful information about RNA structure but currently involve ad hoc assumptions in data processing. We show that simple dilutions, referencing standards (GAGUA hairpins), and HiTRACE/MAPseeker analysis allow rigorous overmodification correction, background subtraction, and normalization for electrophoretic data and a ligation bias correction needed for accurate deep sequencing data. Comparisons across six noncoding RNAs stringently test the proposed standardization of dimethyl sulfate (DMS), 2′-OH acylation (SHAPE), and carbodiimide measurements. Identification of new signatures for extrahelical bulges and DMS "hot spot" pockets (including tRNA A58, methylated *in vivo*) illustrates the utility and necessity of standardization for quantitative RNA mapping.

Structure mapping, also known as footprinting, provides a rapid means for probing nucleic acid conformation at single-nucleotide resolution. New modification chemistries, higher-throughput readouts, multidimensional expansions, error analysis, and resources for sharing data are advancing the approach.[1] Despite powerful insights from separate data sets, ad hoc choices in data processing have precluded robust comparison of chemical reactivities across RNAs and readouts.[2−7] For example, "hot spots" that might signal specific noncanonical features[6,7] in one RNA cannot be confidently established in other RNAs without universal reactivity scales, analogous to problems in nuclear magnetic resonance chemical shift analysis prior to the adoption of referencing samples.[8]

In principle, establishing reactivities should be unambiguous. Modification fractions $r_i$ of nucleotides $i$ can be directly computed from the numbers of "raw" observed products $F_i$ by

$$r_i = \frac{F_i}{F_0 + F_1 + \ldots F_i} \quad (1)$$

(derivation in the Supporting Information). While $F_0$, the number of "full-length" products without chemical modification, is visible for RNA domains of up to 500 nucleotides, accurate quantitation is typically precluded by detector saturation of this strong band in electrophoresis data or by ligation biases in deep sequencing data. Our lab's previous likelihood framework for $F_0$ depended on *a priori* reactivity distributions that were approximate.[2] Aviran et al. explored setting $F_0$ to zero when it could not be measured,[5] a poor assumption under typical "single-hit" conditions. Karabiber et al. proposed equalizing reactivities observed in the 5′ half versus the 3′ half of the data,[3,4] a generally inaccurate approximation. Several recent studies have not applied eq 1.[9] Further complicating cross-experiment comparisons are differences in whether eq 1 is applied to no-modifier control samples, in sequence alignment tools, in error estimation, and in normalization procedures,[2,3,5] as well as a lack of validation protocols.

To address these issues, we implemented two straightforward standardization strategies: (1) dilution comparisons to mitigate saturation and (2) use of universal internal controls (Figure 1A,B). To illustrate, Figure 1C gives capillary electrophoresis (CE) data of primer extension products for the P4−P6 domain of the *Tetrahymena* ribozyme probed with dimethyl sulfate (DMS) to methylate exposed N1/N3 atoms of A/C nucleotides.[10] The saturated peak shape for the fully extended product is apparent; 10-fold dilution of the same sample gave a weaker signal-to-noise ratio overall but an unsaturated, Gaussian shape for the $F_0$ peak (Figure 1D; further dilutions verified the lack of saturation). Automated scaling of these dilution data allowed unbiased measurement of $F_0$ (Figure 1E,F). Application of eq 1, background subtraction, and normalization (see below) gave the reactivity profile in Figure 1F. The final results agreed within error with averaged data collected by different experimenters (Figure 1F and Methods and Figure 1 of the Supporting Information). Further, as expected (but not assumed), DMS reactivities at G and U nucleotides were within error of zero. Tests comparing data from 8-fold variations of DMS and reagents 1-cyclohexyl(2-morpholinoethyl)carbodiimide metho-*p*-toluene sulfonate (CMCT, modifying G/U)[10] and 1-methyl-7-*N*-isatoic anhydride (1M7, modifying 2′-OH; SHAPE[3,4]) further confirmed this standardization (Figure 2 of the Supporting Information).

Independent validation of this procedure came from incorporating "reference" hairpins in 5′ and 3′ flanking cassettes.[3,4] GAGUA hairpin loops (Figure 2a) give strong signals for DMS (at the A's), CMCT (at the bulge U), and 1M7 (all five residues). "Raw" $F_i$ counts were 5-fold lower at the 5′ GAGUA than at the 3′ GAGUA (red bars in Figure 1E), as reverse transcriptases encountered stops in between those

**Received:** March 19, 2014
**Revised:** April 13, 2014
**Published:** April 28, 2014

   dx.doi.org/10.1021/bi5003426 | *Biochemistry* 2014, 53, 3063−3065

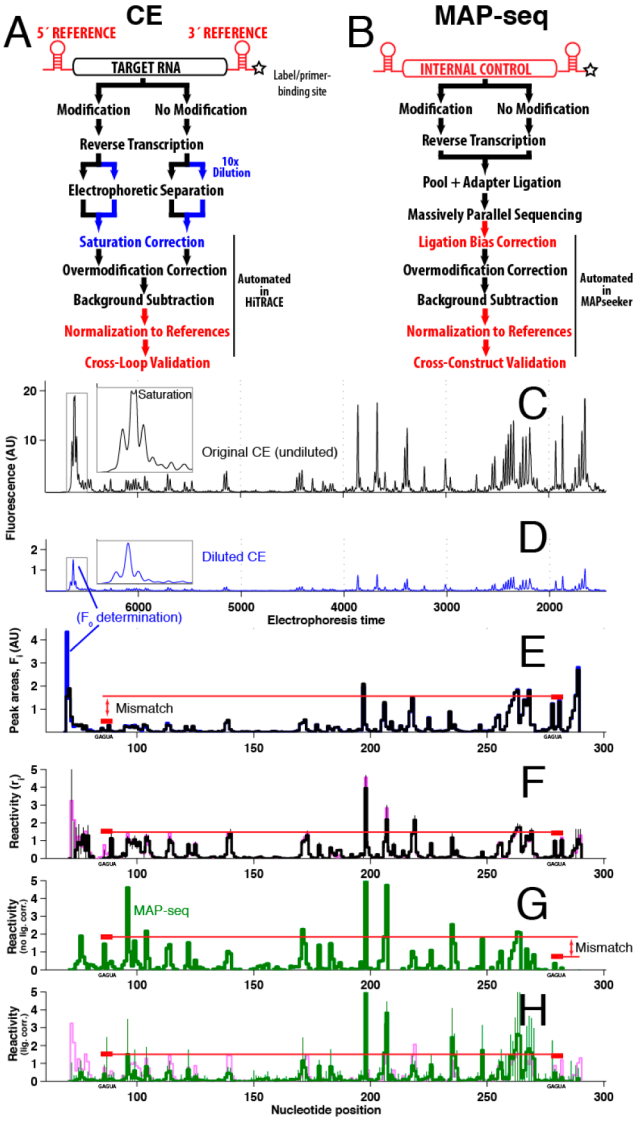



**Figure 1.** Proposed steps to standardize chemical mapping experiments (red and blue text) read out by (A) capillary electrophoresis and (B) deep sequencing (MAP-seq). CE profiles for the P4−P6−2HP RNA probed with DMS at (C) standard dilution and (D) 10-fold dilution. (E) Automated scaling matches diluted sample data to undiluted data. (F) Final reactivity profile (black), validated by data taken at 4-fold lower DMS concentrations (green, nearly indistinguishable) and equality at GAGUA referencing hairpins (red). MAP-seq data for P4−P6 RNA without (F) and with (G) ligation bias correction determined from internal referencing. (H) Overlay of CE and MAP-seq data; errors are standard deviations of replicates (Figure 1 of the Supporting Information).

segments ("overmodification", also called attenuation or signal decay). The equality of the GAGUA final reactivities $r_i$ confirmed accurate overmodification correction and background subtraction of these data (red bars in Figure 1F) and supported use of the GAGUA data as normalization standards.

An alternative readout, MAP-seq (multiplexed accessibility probing), follows nucleic acid modification and primer extension with ligation of an Illumina adapter and deep sequencing, without bias-introducing polymerase chain reaction amplification (Methods of the Supporting Information).[11] We previously observed (through CE) that ligation yields were systematically low for full-length cDNA products. This effect

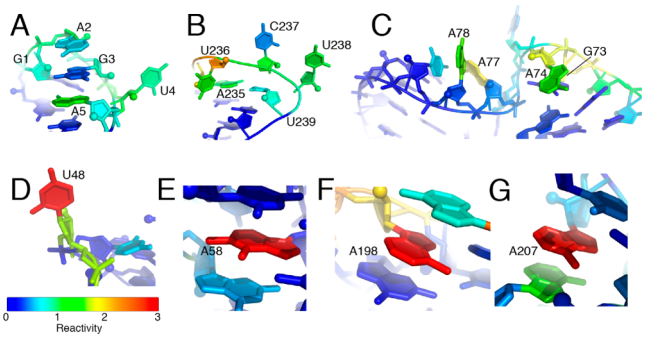



**Figure 2.** Three-dimensional environments associated with high chemical reactivity to Watson−Crick edge modifiers [DMS for A/C and CMCT for G/U (base color)] and/or 2′-OH acylation [1M7 (backbone color)]. (A) GAGUA hairpin sets the normalization scale for DMS (A2 and A5), CMCT (U4), and 1M7 (all nucleotides). (B) L6b from the P4−P6 domain. (C) Interdomain linker from the glycine riboswitch. (D) Bulge in the ligand binding pocket of the adenine riboswitch. (E−G) Pockets promoting high adenosine N1 reactivity and low 2′-OH reactivity in tRNA (N1-methyl shown) (E) and the P4−P6 domain (F and G). Hot spot nucleotides are labeled in panels B−G. Protein Data Bank entries are listed in Table 1 of the Supporting Information.

led to underestimation of $F_0$ and to an apparent discordance between the 5′ and 3′ GAGUA references (red bars, Figure 1G). Nevertheless, the requirement of equality at these sequences allowed automated estimation of a ligation bias correction factor [0.18 in this case (Methods of the Supporting Information)]. Despite involving rather different protocols, the CE and MAP-seq results then agreed within errors estimated from replicates (Figure 1H, and see below).

To comprehensively test the standardization protocol, we took measurements with DMS, CMCT, and 1M7, using both CE and MAP-seq protocols on several structured RNAs, including ligand-bound riboswitches and rRNA domains (Figures 3−8 of the Supporting Information).[2,10] In the MAP-seq experiment, data for the P4−P6−2HP domain established the ligation bias correction factor and normalization for the coloaded RNAs. The agreement within error between reactivities at GAGUA reference hairpins across all constructs and general agreement between CE and MAP-seq data sets confirmed the accuracy of the proposed standardization (Figure 1 of the Supporting Information). No length bias was detected for MAP-seq, but a residual sequence bias was seen in reactive purine-rich segments; these mostly occurred in flanking sequences outside the structured RNA domains (Figures 3−8 of the Supporting Information). In both CE and MAP-seq data, normalization to GAGUA references exposed limitations of prior heuristics that normalize based on high percentile values within each RNA (or in 5′ and 3′ halves);[2−4,9,10] these values in fact vary by >2-fold across the different RNAs.

The standardization procedures allowed the identification of 33 hot spot nucleotides, defined here as those giving DMS, CMCT, or 1M7 reactivity of >1.5, well above control values (1.0) established by GAGUA references (Table 2 of the Supporting Information). First, in agreement with conventional use of these data to infer secondary structure,[10] all 16 cases of high DMS/CMCT/1M7 reactivities observed within stretches of more than two residues corresponded to apical loops (Figure 2B) or unpaired "linkers" (Figure 2C). Second, three isolated adenosines with high 1M7 but low DMS reactivity were stacked on one face, a structural feature previously requiring differential


dx.doi.org/10.1021/bi5003426 | Biochemistry 2014, 53, 3063−3065

SHAPE measurements for identification.[6] Third, all seven isolated highly CMCT/1M7-reactive uridines and two highly 1M7-reactive adenosines were extrahelical bulges[7] (Figure 2D), a powerful signature for guiding or validating tertiary structure modeling.[12] Most intriguing were five adenosines with DMS reactivities of >1.5 but negligible 1M7 reactivity (Figure 2E–G). Each of these adenosines showed Hoogsteen edge burial and nucleobase stacking on both faces; such burial information should be useful in tertiary structure modeling. The most DMS-reactive nucleotide, A58 in *Saccharomyces cerevisiae* tRNA(phe) (Figure 2E), is also methylated at the N1 position *in vivo*.[13] The pocket around DMS hot spot nucleotides may thus be under selection for electronegativity to enhance enzymatic reaction or hydrogen bonding to partners. As further examples, A198 and A207 (Figure 2F,G) in the isolated P4–P6 domain are buried, but N1 atoms are available for contacts in the full *Tetrahymena* ribozyme or recognition by protein partners. These signatures could not be identified unambiguously in prior work because of uncertain data scaling.

The inclusion of dilution samples and referencing hairpins allows standardization, validation, and deeper analysis of structure mapping experiments at negligible additional cost. For CE studies, obtaining the necessary data simply involves diluting the prepared samples into running buffer and repeating electrophoresis and HiTRACE/HiTRACE-Web analysis[14] (Figure 1A). Inclusion of GAGUA hairpins was used here to test the overmodification correction and normalize CE data but was only strictly necessary in MAP-seq experiments. In fact, just a single construct with flanking reference hairpins needs to be doped into the MAP-seq RNA pool; standardization is then automated via MAPseeker analysis[11] (Figure 1B). The general adoption of simple standardization steps, and their extension to very long transcripts and to other solution conditions and modifiers, should help RNA structure mapping data become more accurate and more transferrable between molecules and experiments.

## ■ ASSOCIATED CONTENT

### Ⓢ Supporting Information

Derivation of eq 1, experimental methods, CE/MAP-seq comparisons, and a table of sequences, Protein Data Bank entries, and RNA Mapping Database entries for deposited data. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION


### Corresponding Author

*E-mail: rhiju@stanford.edu. Phone: (650) 723-5976. Fax: (650) 723-6783.

### Author Contributions

W.K. and T.H.M. contributed equally to this work.



### Funding

Work was funded by the Burroughs-Wellcome Foundation (CASI 1007236.01 to R.D.), a Stanford Graduate Fellowship (S.T.), National Research Foundation of Korea (No. 2011-0009963 to S.Y.), the National Institutes of Health (5T32GM007276 to T.H.M. and R01GM102519 to R.D.), and the Keck Foundation.


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS


We thank P. Cordero, C. Cheng, B. Stoner, and Das lab members for helpful discussions and S. Mortimer and F. V. Cochran for assistance with 1M7 synthesis.


## ■ REFERENCES


(1) Peattie, D. A., and Gilbert, W. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 4679–4682. Regulski, E. E., and Breaker, R. R. (2008) *Methods Mol. Biol. 419*, 53–67. Kladwang, W., and Das, R. (2010) *Biochemistry 49*, 7414–7416. Rocca-Serra, P., Bellaousov, S., Birmingham, A., Chen, C., Cordero, P., Das, R., Davis-Neulander, L., Duncan, C. D., Halvorsen, M., Knight, R., Leontis, N. B., Mathews, D. H., Ritz, J., Stombaugh, J., Weeks, K. M., Zirbel, C. L., and Laederach, A. (2011) *RNA 17*, 1204–1212. Cordero, P., Lucks, J. B., and Das, R. (2012) *Bioinformatics 28*, 3006–3008.

(2) Kladwang, W., Vanlang, C. C., Cordero, P., and Das, R. (2011) *Biochemistry 50*, 8049–8056.

(3) Karabiber, F., McGinnis, J. L., Favorov, O. V., and Weeks, K. M. (2013) *RNA 19*, 63–73.

(4) Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M. (2005) *J. Am. Chem. Soc. 127*, 4223–4231.

(5) Aviran, S., Lucks, J. B., and Pachter, L. (2011) *Proceedings of the 49th Allerton Conference on Communication, Control, and Computing*, 1743–1750.

(6) Steen, K. A., Rice, G. M., and Weeks, K. M. (2012) *J. Am. Chem. Soc. 134*, 13160–13163.

(7) McGinnis, J. L., Dunkle, J. A., Cate, J. H., and Weeks, K. M. (2012) *J. Am. Chem. Soc. 134*, 6617–6624.

(8) Aeschbacher, T., Schubert, M., and Allain, F. H. (2012) *J. Biomol. NMR 52*, 179–190.

(9) Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010) *Nature 467*, 103–107. Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., and Assmann, S. M. (2014) *Nature 505*, 696–700. Kwok, C. K., Ding, Y., Tang, Y., Assmann, S. M., and Bevilacqua, P. C. (2013) *Nat. Commun. 4*, 2971. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J. S. (2014) *Nature 505*, 701–705.

(10) Cordero, P., Kladwang, W., VanLang, C. C., and Das, R. (2012) *Biochemistry 51*, 7037–7039.

(11) Seetin, M. G., Kladwang, W., Bida, J. P., and Das, R. (2014) *Methods Mol. Biol. 1086*, 95–117.

(12) Sripakdeevong, P., Kladwang, W., and Das, R. (2011) *Proc. Natl. Acad. Sci. U.S.A. 108*, 20573–20578.

(13) Sengupta, R., Vainauskas, S., Yarian, C., Sochacka, E., Malkiewicz, A., Guenther, R. H., Koshlap, K. M., and Agris, P. F. (2000) *Nucleic Acids Res. 28*, 1374–1380.

(14) Yoon, S., Kim, J., Hum, J., Kim, H., Park, S., Kladwang, W., and Das, R. (2011) *Bioinformatics 27*, 1798–1805. Kim, H., Cordero, P., Das, R., and Yoon, S. (2013) *Nucleic Acids Res. 41*, W492–W498.


## ■ NOTE ADDED AFTER ASAP PUBLICATION

This Rapid Report was published ASAP on May 7, 2014. Reference 5 has been updated and the corrected version was reposted on May 8, 2014.

# Supporting Information for "Standardization of RNA chemical mapping experiments"

Wipapat Kladwang[1], Thomas H. Mann[1], Alex Becka[1], Siqi Tian[1], Hanjoo Kim[2], Sungroh Yoon[2], and Rhiju Das[1,3,†]

[1]Department of Biochemistry, Stanford University, Stanford CA 94305
[2]Department of Electrical and Computer Engineering, Seoul National University, Seoul 151-744, Korea
[3]Department of Physics, Stanford University, Stanford CA 94305

**This Supporting Information document contains the following sections:**

**SUPPORTING DERIVATION**

*Relation between observed and actual product fractions*

Determining chemical reactivity profiles for nucleic acids requires taking into account how a chemical modification internal to a fragment can lower the probability of observing longer fragments. For completeness, we derive the relation between observed and actual product fractions [main text eq. (1)] here. In the case of reverse transcription, let the first reverse transcribed nucleotide be $N$ (i.e., the maximum product length), the total number of products be $M$, and the fraction of chemical modified nucleotides at each position $i = 1, \ldots N$ be $r_i$. Then, the number of full length products $F_0$, corresponding to events with no modification at any internal site is:

$$F_0 = \left(1 - r_1\right)\left(1 - r_2\right)\ldots\left(1 - r_N\right)M \,, \qquad \text{(S1)}$$

The number of products $F_i$ corresponding to modification at each nucleotide $i$ are:

$$F_i = r_i\left(1 - r_{i+1}\right)\left(1 - r_{i+2}\right)\ldots\left(1 - r_N\right)M \qquad \text{(S2)}$$

See also refs[1,2]. Partial summation of these values from the 5′ end gives:

$$\begin{aligned} F_0 + F_1 &= \left(1 - r_2\right)\left(1 - r_3\right)\ldots\left(1 - r_N\right)M \\ F_0 + F_1 + F_2 &= \left(1 - r_3\right)\ldots\left(1 - r_N\right)M \\ F_0 + F_1 + \ldots F_N &= M \end{aligned} \qquad \text{(S3)}$$

Combining eqs. (S2) and (S3) gives the sought relation of $r_i$ to the observed $F_i$:

$$r_i = \frac{F_i}{F_0 + F_1 + \ldots F_i} \,,$$

The equation corresponds to the fraction of reverse transcriptases that stopped at position $i$, compared the total number of reverse transcriptases that reached position $i$ (and either stopped or proceeded beyond). This derivation also holds for protocols that involve chemical or enzymatic cleavage of end-labeled nucleic acids instead of primer extension, with $N$ marking the position of the 3′-end label. For 5′-end labels, the same relations hold but with indices $i$ reversed in order; in all cases $F_0$ should correspond to unmodified nucleic acid. The above values $r_i$ range from 0 to 1. The number of modification events $R_i$ is a better estimate of chemical reactivity which scales linearly with modifier

concentration; it can be estimated from the relation $r_i = 1 - \exp(-R_i)$. However, $R_i = r_i$ for $r_i << 1$, as was the case herein. Last, background subtraction and scaling based on internal standards (see main text) gives $r_i^{norm}$, which can exceed 1.

## EXPERIMENTAL METHODS

### Data and software availability

All analysis steps have been implemented in two freely available software packages HiTRACE (for capillary electrophoresis analysis; http://www.hitrace.org for HiTRACE-Web server or MATLAB software download) and MAPseeker (for MAP-seq deep sequencing analysis; https://github.com/MAPseeker for software download). See below for description of processing steps and implementations. Data have been deposited in the RNA Mapping Database[3] (http://rmdb.stanford.edu) under accession codes listed in SI Table 1.

### Preparation of RNA

RNA preparation procedures followed those in experiments described previously[4,5], with small modifications noted here. Briefly, DNA templates were produced through PCR assembly of oligonucleotides of length 60 nucleotides or smaller (Integrated DNA Technologies) using Phusion polymerase (Finnzymes, MA). DNA templates were designed with the T7 RNA polymerase promoter (TTCTAATACGACTCACTATA) at their 5′ ends. A custom reverse transcription primer-binding site (AAAGAAACAACAACAACAAC) was included at the 3′ terminus of each template. See Table 1. Flanking sequences, here including referencing hairpin stems, were screened computationally to not interact with the target RNA sequence on the NUPACK server.[6] (Data were consistent within error and scaling with prior measurements with different flanking sequences.[2,4,7]) RNA transcribed with T7 RNA polymerase (New England Biolabs) was purified using the RNA Clean & Concentrator 5 kit (Zymo Research).

### Chemical mapping experiments read out by capillary electrophoresis

Chemical mapping procedures with capillary electrophoresis (CE) followed those in these experiments described previously.[4,5] Briefly, modification reactions were performed in 20 μL reactions containing 1.2 pmol RNA, 50 mM Na-HEPES (pH 8.0), and 10 mM $MgCl_2$. Ligand-binding RNAs were incubated with specified ligands at room temperature for 30 minutes prior to mapping. Chemical probes were used at the following final concentrations: DMS (0.125% v/v for P4P6-2HP, varied from 0.03125% to 0.5% where noted; added with 0.25% ethanol), CMCT in water (2.6 mg/mL standard; 0.66–10.5 mg/mL where noted), 1M7 (1.05 mg/mL standard; 0.2625–4.2 mg / mL where noted; stock prepared in anhydrous DMSO gave final DMSO concentration of 25%). Chemical probes were allowed to react for 15 minutes prior to quenching. The reaction quench for 1M7 and CMCT contained 5.0 μL of 0.5 M Na-MES (pH 6.0), 3 μL of 3 M NaCl, 1.5 μL of oligo-dT beads (poly(A) purist, Ambion), and 0.25 μL of a 0.25 μM 5′-rhodamine-green labeled primer (Table 1), complementary to the reverse transcription primer-binding site at the RNA 3′ ends. The reaction quench for DMS was identical except that the Na-MES component was replaced with 5 μL of 2-mercaptoethanol. This quench mixture allowed for purification and reverse transcription on magnetic beads. Chemically modified RNAs were reverse transcribed with Superscript III Reverse Transcriptase (Life Technologies). RNA was subsequently hydrolyzed for 3 minutes at 90 °C in 0.2 M NaOH. After pH neutralization and ethanol rinsing, cDNAs were eluted into 10 μL HiDi formamide (Life Technologies) and co-loaded with a ROX-350 standard ladder (Life Technologies) for electrophoresis on ABI 3130 or 3700 sequencers.

CE data were quantitated with HiTRACE[8] to give observed product frequencies $F_i^{\text{observed}}$. Fitting errors were estimated with the function `fit_to_gaussians` based on analytical computation of the sum of the squares of standard deviations of $F_i^{\text{observed}}$ upon shifting all band positions by ±0.5 of the mean band-to-band spacing. Reactions without chemical modification gave estimates of backgrounds, $F_i^{\text{background}}$. We note that the units of these measurements are arbitrary; final reactivity fractions $r_i$ are calculated based on ratios of these observed products to sums of products measured within the same trace (see Supporting Derivation and next section).

**Quantitative analysis of nucleotide reactivities (capillary electrophoresis)**

Four data processing steps were carried out for analysis of chemical mapping experiments (see also main text Fig. 1a) from capillary electrophoresis experiments. The entire pipeline is available through an online workflow on the HiTRACE-Web server as well as from the MATLAB implementation via a single HiTRACE script `get_reactivities`.

1. Saturation correction from dilution samples

Applying eq. (1) of the main text (see also SI Appendix) requires accurate quantitation of observed products $F_i$. In most electrophoresis experiments, the full-length band ($F_0$) and occasionally other strong bands saturate the detector. In addition to our fully concentrated samples, we acquired capillary electrophoretic traces for 10-fold dilutions of each sample, by removing 1 μL of the reactions prepared for ABI sequencers into running buffer of 9 μL of HiDi formamide with the ROX-350 standard. For some experiments, it may be necessary to dilute further than 10-fold, although carrying out this additional dilution series was not necessary for the the linearity range of our CE detector and maximum band intensity observed in our experiments.

A scalefactor $\alpha$ was determined to match the resulting $F_i^{\text{diluted}}$ to the original $F_i^{\text{observed}}$,

$$\alpha = \frac{\sum_{i=\text{non-saturated}} F_i^{\text{observed}}}{\sum_{i=\text{non-saturated}} F_i^{\text{diluted}}}$$

where "non saturated" refers to the subset of nucleotides at which the diluted data matched the undiluted data. This subset was determined by beginning with the full set of measured nucleotides, determining $\alpha$, computing residuals $\delta_i = F_i^{\text{diluted}} - F_i^{\text{observed}}$, filtering out any nucleotides $i$ at which $\delta_i$ exceeded the mean of $\delta_i$ by more than 1.5 times the standard deviation of $\delta_i$; and iterating the procedure a total of three times. For each nucleotide $i$ that was filtered out as 'saturated' by this procedure, $F_i^{\text{observed}}$ and its estimated error was replaced with $\alpha F_i^{\text{diluted}}$ and its error (scaled by $\alpha$). The same procedure was used to correct for saturation in control measurements without chemical

modification $F_i^{\text{background}}$. The procedure, automated in the script `unsaturate` in HiTRACE, returns an image showing saturated residues.

2. Over-modification correction

Main text equation 1 (see also Supporting Derivation) was applied to transform the (saturation corrected) observed product values $F_i^{\text{observed}}$ and $F_i^{\text{background}}$ to give modification fractions $r_i^{\text{observed}}$ and $r_i^{\text{background}}$. Relative errors on $r_i^{\text{observed}}$ were taken from errors on $F_i^{\text{observed}}$. [Additional relative errors due to summation across the denominator $F_0 + F_1 + \ldots F_i$ would give rise to correlated errors across the entire profile and were not modeled here due to their complexity and to their generally small contributions.] This procedure was automated in the script `correct_for_attenuation`.

3. Background subtraction

The modification fraction $r_i$ at each nucleotide $i$ due to chemical modification was given by $r_i^{\text{observed}} - r_i^{\text{background}}$, with error estimated by summing errors of components in quadrature. This procedure was automated in the script `subtract_array`.

4. Normalization to referencing segments

Inclusion of at least one referencing hairpin with pentaloop sequence GAGUA enabled normalization, giving final values $r_i^{\text{norm}}$ that were independent of the chosen modifier concentration and time. Modification fractions $r_i$ were scaled so that the underlined nucleotides gave mean reactivities of 1.0: G<u>A</u>GU<u>A</u> (DMS), GAG<u>U</u>A (CMCT), and <u>GAGUA</u> (SHAPE). In constructs with two hairpins in both 5′- and 3′- flanking sequences, we used data for the 3′- GAGUA hairpin for normalization, due to high errors in background subtraction for GAGUA hairpins in 5′ flanking sequences. This procedure was automated in the script `apply_normalization`.

All steps above are included in HiTRACE software, and a step-by-step workflow is available as a default stage in the online HiTRACE-web server.[9]

**Chemical mapping experiments read out by deep sequencing (MAP-seq)**

The detailed MAP-seq protocol has been presented in ref.[10] and is briefly summarized here. Chemical modification reactions were carried out as in CE reactions in 10 mM $MgCl_2$ and 50 mM Na-HEPES, pH 8.0, but with 4-8 pmols of each RNA in 50 μL volumes. In one set of experiments, the six RNAs, along with other RNAs (including ligand-binding domains for an adenosyl-cobalamin and flavin mononucleotide riboswitch; data not shown), were subjected to modification at different concentrations of DMS (0.125% and 0.5% v/v final), 1M7 (1.05 and 4.24 mg/mL final), or no reagent, either without added ligand or with a ligand mixture (final concentrations of 5 mM adenine, 10 μM cyclic-diguanosine monophosphate, 10 mM glycine, 200 μM flavin mononucleotide, and adenosyl cobalamin, 70 μM) in 12 samples. In a second set of experiments, each of the six RNAs was modified by DMS (0.125% v/v final), 1M7 (1.05 mg/mL final), or no reagent, with ligand (5 mM adenine, 10 μM cidGMP, or 10 mM glycine for the relevant riboswitches; a mixture of all three for the other RNAs) or no ligand, in 36 separate samples. In each case, modified RNA in 50 μL volumes was precipitated by addition of 10 μL 3 M Na-acetate and 330 μL cold ethanol; microcentrifugation; washed with 70 % ethanol; and resuspended in 9.7 μL deionized water. Reverse transcription reactions were carried out as in CE experiments except scaled up to 15 μL volumes and using 5′-FAM-labeled Illumina-Oligo C-containing primers with 12 nt barcodes; pulldowns were carried out with DynaBeads (Life Technologies) displaying the reverse complement to Oligo C (Oligo C′, TGTGTAGATCTCGGTGGTCGCCGTATCATTTTTTTTTTTT-3′-double-biotin) and cDNA was resuspended (with beads remaining) into 2 μL deionized water. Aliquots of these reactions were run by CE to confirm reverse transcription; for the second set of experiments, each sample involved a single RNA, and so these data could be analyzed by HiTRACE and gave data consistent with CE measurements above. To complete the MAP-seq protocol, sets of four samples were pooled into 8 μL volumes and cDNAs were ligated to the second Illumina adapter (1.25 μM 5′-phosphate-

AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTC TTCTGCTTG-3′-phosphate) with CircLigase I (5 U μL) in 2.5 mM $MnCl_2$, 1x CircLigase I buffer, 4% PEG 1500, and 50 μM ATP for 2 hours in 50 μL volumes at 68 °C, followed by 10 minutes at 80 °C for deactivation of the ligase. Samples were again pulled down by magnetic separation and washed, and ligated cDNA concentrations were quantitated by loading on ABI 3130 sequencer with FAM-labeled standards. Samples were eluted by addition of 4.5 μL 10 mM EDTA, 95% formamide, heating to 90 °C for 2 minutes, cooling to room temperature; and addition of 0.5 μL 5 nM PhiX control (Illumina). The sample was then prepared for loading onto Miseq v2 kits following manufacturer instructions.

Data processing was carried out with the `quick_look_mapseeker` routine in MAPseeker software, as described in ref.[10]. All MAPseeker steps requiring an internal control are automatically activated when the P4P6-2HP sequence with flanking GAGUA hairpins (SI Table 1) is included in the run and specified as one of the probed RNAs in `RNA_sequences.fasta`. In particular, the critical ligation bias correction term for the fully extended product (relative to the average sequence) was determined by applying an initial value of $\beta = 1.0$ to find the background-subtracted reactivities of the control P4P6-2HP sequence; testing for equality at GAGUA sequences at the 5′ and 3′ ends; and then optimizing $\beta$ to give exact equality at the GAGUA sequences by numerical search with MATLAB's `fminbnd` function. Over-modification correction, background subtraction, and normalization to GAGUA segments were carried out as described above for C.E.; these steps occur automatically in the MAPseeker workflow.

We observed that normalization of MAP-seq data at flanking GAGUA hairpins produced reactivities that were higher than CE measurements within the target of interest, presumably due to systematic ligation bias at those nucleotides relative to other sequences. For Fig. 1 and SI Figures 3-8, we therefore used the average MAP-seq reactivity value over nucleotides in the P4P6-2HP RNA, compared to their average value in CE data, to provide correction factors applied to all other MAP-seq data (0.55 and 0.40 for 1M7 and DMS, respectively). Note that for any future MAP-seq experiments

with the P4P6-2HP, the normalization scale will be automatically set by MAPseeker based on GAGUA hairpins, but will need to be re-scaled by the above factors if matching to CE experiments is required.

**Averaging and error estimations based on multiple replicates**

For CE experiments, we found that estimates of error due to peak fitting or from standard deviations within each experiment generally underestimated errors estimated from repeating experiments. For MAP-seeker experiments, estimates of error based on Poisson counting statistics also gave underestimates of error for MAP-seq runs with >100,000 counts for each RNA (this high-statistics limit was the case herein). For example, a previous averaging procedure, which estimated final reactivity errors based solely on propagating estimates from CE peak-fitting, did not capture the high tRNA(phe) DMS reactivity at A58 (SI Fig. 3).[2,7] A user error assigned this strong band to an adjacent residue in some replicates; the deviation between the replicates should have been reflected in high errors at both positions, but the propagated CE errors, dominated by replicates with zero reactivity, showed low reactivities. The identification of this issue led to developments of an automated sequence assignment tool (available in HiTRACE and HiTRACE-web; SRY, HY, RD, in prep.), a visual display of averaging results, and a more conservative error estimation procedure.

The new averaging and error estimation procedure is encoded in the HiTRACE function `average_data_filter_outliers` (and a wrapper function for data formatted in the RDAT format[3], `rdat_combine`). At each nucleotide position, the reactivity values from different measurements were averaged, weighted by the inverse of errors estimated from peak fitting (as is returned by HiTRACE CE fits) or Poisson error (as is returned by MAPseeker error). (To avoid extremely low error points from dominating this average, the error on each input measurements was set to 10% of the reactivity if it was estimated t be lower.) The standard error on this average value was taken as the standard deviation among measurements, divided by the square root of the number of observations. If any of the measurements gave a value more than five standard deviations from the original average, that entire measurement was automatically flagged as an outlier, and the average

was recalculated without this value. In rare cases, an entire replicate gave a mean discrepancy at all nucleotides more than 2.5 standard deviations from the replicate average; this measurement was also flagged as an outlier and not included in the final calculation. All measurements and averaging are displayed graphically in both heat-map and trace overlay formats to allow visual assessment of sequence assignents, automatically assigned outliers, and other variability across measurements. An example of this display of measurement averaging and error estimation is given in SI Fig. S1.

**Data visualization**

Data figures prepared in MATLAB (http://www.mathworks.com) and PyMol (http://www.pymol.org). 3D coloring scripts are freely available at https://github.com/DasLab/pymol_daslab.

## SUPPORTING REFERENCES


1. Aviran, S., Trapnell, C., Lucks, J. B., Mortimer, S. A., Luo, S., Schroth, G. P., Doudna, J. A., Arkin, A. P., and Pachter, L. (2011), *Proc Natl Acad Sci U S A 108*, 11069-11074.
2. Kladwang, W., Vanlang, C. C., Cordero, P., and Das, R. (2011), *Biochemistry 50*, 8049-8056.
3. Cordero, P., Lucks, J. B., and Das, R. (2012), *Bioinformatics 28*, 3006-3008.
4. Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011), *Nat Chem 3*, 954-962.
5. Seetin, M. G., Kladwang, W., Bida, J. P., and Das, R. (2013) Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol, In *RNA Folding (Methods in Molecular Biology)* (Waldsich, C., Ed.), p in press.
6. Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., Pierce, M. B., Khan, A. R., Dirks, R. M., and Pierce, N. A. (2011), *Journal of computational chemistry 32*, 170-173.
7. Cordero, P., Kladwang, W., VanLang, C. C., and Das, R. (2012), *Biochemistry 51*, 7037-7039.
8. Yoon, S., Kim, J., Hum, J., Kim, H., Park, S., Kladwang, W., and Das, R. (2011), *Bioinformatics 27*, 1798-1805.
9. Kim, H., Cordero, P., Das, R., and Yoon, S. (2013), *Nucleic Acids Res 41*, W492-498.
10. Seetin, M. G., Kladwang, W., Bida, J. P., and Das, R. (2014), *Methods Mol Biol 1086*, 95-117.

**SUPPORTING TABLE 1. Nucleic acid sequences and database IDs**. Sequences written from 5′ to 3′. Structured RNA domain of interest highlighted in blue. Reference GAGUA hairpins highlighted in red; some sequences have two references to test overmodification correction. Protein DataBank IDs give crystallographic models used to assess 3D environments, and accession IDs of data collected herein and deposited in the RNA Mapping Database are also given.

| **Molecule** | **Sequence** (conventional numbering) | **PDB** | **RMDB IDs**[a] |
|---|---|---|---|
| **tRNA-1HP**, umodified tRNA (phe), S. cerevisiae | GGAACAAACAAAACAGCGGAUUUAGCU CAGUUGGGAGAGCGCCAGACUGAAGAU CUGGAGGUCCUGUGUUCGAUCCACAGA AUUCGCACCAAAACGUUAAGGAGUACU UAACCAAAGAAACAACAACAACAAC (–15 to 117) | 1EHZ | TRNAPH_DMS_0005<br>TRNAPH_CMC_0005<br>TRNAPH_1M7_0005<br>TRNAPH_DMS_0006<br>TRNAPH_CMC_0006<br>TRNAPH_1M7_0006 |
| **ADD-2HP**, adenine riboswitch, V. vulnificus | GGAAAGCAAUUCGAGUAGAAUUGGAAA GGGAAAGAAACGCUUCAUAUAAUCCUA AUGAUAUGGUUUGGGAGUUUCUACCAA GAGCCUUAAACUCUUGAUUAUGAAGUG AAAACAAAGUUAAGGAGUACUUAACAC AAAGAAACAACAACAACAAC (–25 to 129) | 1Y26 | ADDRSW_DMS_0005<br>ADDRSW_CMC_0005<br>ADDRSW_1M7_0005<br>ADDRSW_DMS_0006<br>ADDRSW_CMC_0006<br>ADDRSW_1M7_0006 |
| **cidGMP-2HP**, Cyclic di-GMP riboswitch, V. cholerae | GGAAAAAUGUCACGCACAGGGCAAACC AUUCGAAAGAGUGGGACGCAAAGCCUC CGGCCUAAACCAGAAGACAUGGUAGGU AGCGGGGUUACCGAUGGCAAAAUGCAU ACAAACCGUUAAGGAGUACUUAACAAA GAAACAACAACAACAAC (0 to 151) | 3MXH | CDIGMP_DMS_0005<br>CDIGMP_CMC_0005<br>CDIGMP_1M7_0005<br>CDIGMP_DMS_0006<br>CDIGMP_CMC_0006<br>CDIGMP_1M7_0006 |
| **5S-2HP**, 5S rRNA, E. coli | GGAAAGCAAUUCGAGUAGAAUUGGAAA GGGAAAGAAAUGCCUGGCGGCCGUAGC GCGGUGGUCCCACCUGACCCCAUGCCGA ACUCAGAAGUGAAACGCCGUAGCGCCG AUGGUAGUGUGGGGUCUCCCCAUGCGA GAGUAGGGAACUGCCAGGCAUAAAACA GUUAAGGAGUACUUAACAAACAAAGAA ACAACAACAACAAC (–37 to 166) | 3OFC | 5SRRNA_DMS_0005<br>5SRRNA_CMC_0005<br>5SRRNA_1M7_0005<br>5SRRNA_DMS_0006<br>5SRRNA_CMC_0006<br>5SRRNA_1M7_0006 |
| **P4P6-2HP**, P4-P6 domain of *Tetrahymena* ribozyme | GGCCAAAGGCGUCGAGUAGACGCCAAC AACGGAAUUGCGGGAAAGGGGUCAACA GCCGUUCAGUACCAAGUCUCAGGGGAA ACUUUGAGAUGGCCUUGCAAAGGGUAU GGUAAUAAGCUGACGGACAUGGUCCUA ACCACGCAGCCAAGUCCUAAGUCAACA GAUCUUCUGUUGAUAUGGAUGCAGUUC AAAACCAAACCGUCAGCGAGUAGCUGA CAAAAAGAAACAACAACAACAAC (71 to 309) | 1GID | TRP4P6_DMS_0005<br>TRP4P6_CMC_0005<br>TRP4P6_1M7_0005<br>TRP4P6_DMS_0006<br>TRP4P6_CMC_0006<br>TRP4P6_1M7_0006 |
| **FN-2HP**, double glycine riboswitch, F. nucleatum | GGCAAUUCGAGUAGAAUUGACAGAGAG GAUAUGAGGAGAGAUUUCAUUUUAAUG AAACACCGAAGAAGUAAAUCUUUCAGG UAAAAAGGACUCAUAUUGGACGAACCU CUGGAGAGCUUAUCUAAGAGAUAACAC CGAAGGAGCAAAGCUAAUUUUAGCCUA AACUCUCAGGUAAAAGGACGGAGAAAA CACAAGUUCAGGAGUACUGAACCAAAG AAACAACAACAACAAC (–27 to 204) | 3P49 | FNGLYC_DMS_0005<br>FNGLYC_CMC_0005<br>FNGLYC_1M7_0005<br>FNGLYC_DMS_0006<br>FNGLYC_CMC_0006<br>FNGLYC_1M7_0006 |

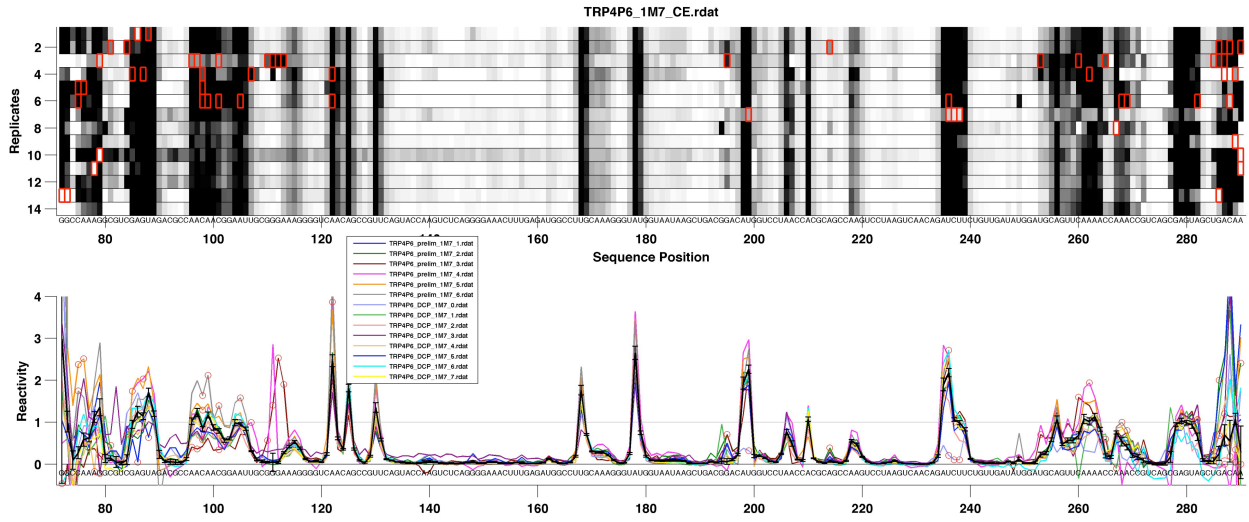
[a]RMDB IDs ending in 0005 store CE data; those ending in 0006 store MAP-seq data.

**SUPPORTING TABLE 2. Hotspot nucleotides.** Nucleotides within structured RNAs that gave DMS, CMCT, or 1M7 reactivity above 1.5 (average from CE measurements).

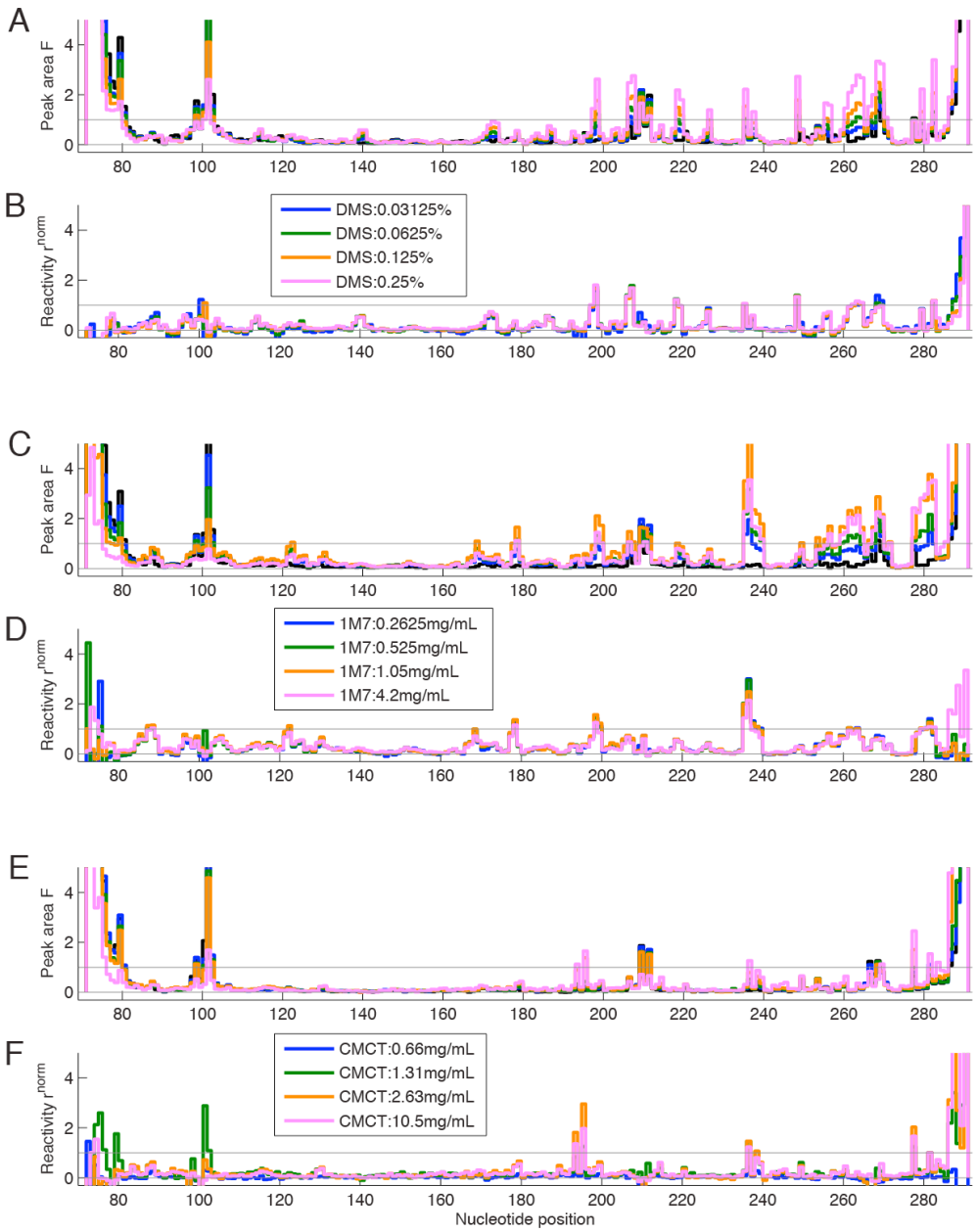
| | | Base reactivity[a] | | 2′-OH reactivity[b] | | |
|---|---|---|---|---|---|---|
| **Molecule** | **Nucleotide** | **value** | **error** | **value** | **error** | **Feature** |
| tRNA(phe) | U16 | **1.54** | 0.12 | 0.52 | 0.03 | extrahelical bulge |
| tRNA(phe) | U17 | **1.81** | 0.18 | 0.55 | 0.05 | extrahelical bulge |
| tRNA(phe) | U33 | 0.43 | 0.04 | **1.69** | 0.13 | apical loop |
| tRNA(phe) | G34 | 0.37 | 0.02 | **2.23** | 0.15 | apical loop |
| tRNA(phe) | A35 | **1.53** | 0.08 | **1.85** | 0.13 | apical loop |
| tRNA(phe) | A36 | **1.21** | 0.05 | **1.63** | 0.08 | apical loop |
| tRNA(phe) | A38 | **1.84** | 0.09 | **1.55** | 0.07 | apical loop |
| tRNA(phe) | A58[c] | **5.34** | 0.5 | 0.1 | 0.03 | buried pocket |
| Adenine riboswitch | U36 | **1.7** | 0.23 | 0.18 | 0.03 | extrahelical bulge |
| Adenine riboswitch | U48 | **3.64** | 0.28 | **1.58** | 0.09 | extrahelical bulge |
| Adenine riboswitch | U62 | **1.61** | 0.27 | 0.87 | 0.01 | extrahelical bulge |
| Cyclic-diGMP riboswitch | A23 | **1.76** | 0.19 | 0.03 | 0.01 | buried pocket |
| Cyclic-diGMP riboswitch | A24 | **2.05** | 0.21 | 0.06 | 0.02 | buried pocket |
| Cyclic-diGMP riboswitch | A68 | 0.97 | 0.11 | **1.78** | 0.12 | apical loop |
| P4-P6 RNA | A122 | 0.78 | 0.05 | **2.46** | 0.14 | 1-stack |
| P4-P6 RNA | A125 | 0.86 | 0.06 | **1.81** | 0.09 | extrahelical bulge |
| P4-P6 RNA | U168 | 0.18 | 0.02 | **1.76** | 0.11 | bulge (makes H-bond) |
| P4-P6 RNA | A178 | 0.69 | 0.06 | **2.65** | 0.16 | 1-stack |
| P4-P6 RNA | A198 | **4.55** | 0.30 | 1.93 | 0.15 | buried pocket |
| P4-P6 RNA | U199 | **1.36** | 0.12 | **2.26** | 0.1 | extrahelical bulge |
| P4-P6 RNA | A207 | **2.83** | 0.33 | 0.49 | 0.06 | buried pocket |
| P4-P6 RNA | A219 | **2.06** | 0.10 | 0.46 | 0.03 | buried pocket |
| P4-P6 RNA | A235 | 1.16 | 0.10 | **1.9** | 0.15 | apical loop |
| P4-P6 RNA | U236 | 1.18 | 0.10 | **2.18** | 0.10 | apical loop |
| Glycine riboswitch | U21 | 0.92 | 0.13 | **2.31** | 0.26 | apical loop |
| Glycine riboswitch | A56 | **1.58** | 0.13 | 0.15 | 0.02 | extrahelical bulge |
| Glycine riboswitch | G73 | 0.11 | 0.02 | **1.84** | 0.08 | linker |
| Glycine riboswitch | A74 | 1.44 | 0.07 | **1.70** | 0.08 | linker |
| Glycine riboswitch | A77 | **1.84** | 0.07 | 0.36 | 0.04 | linker |
| Glycine riboswitch | A78 | **1.52** | 0.13 | 0.27 | 0.03 | linker |
| Glycine riboswitch | U96 | 0.14 | 0.10 | **2.35** | 0.37 | apical loop |
| Glycine riboswitch | A98 | 1.25 | 0.06 | **1.64** | 0.07 | apical loop |
| Glycine riboswitch | A125 | **1.52** | 0.06 | **1.78** | 0.11 | apical loop |

[a]DMS reactivity for A, C; CMCT reactivity for G, U. (1.0 corresponds to A's and U in GAGUA, respectively.)
[b]1M7 reactivity (1.0 corresponds to average reactivity over the five nucleotides of GAGUA).
[c]Not observed in prior work; see SI Methods "Averaging and error estimations based on multiple replicates".

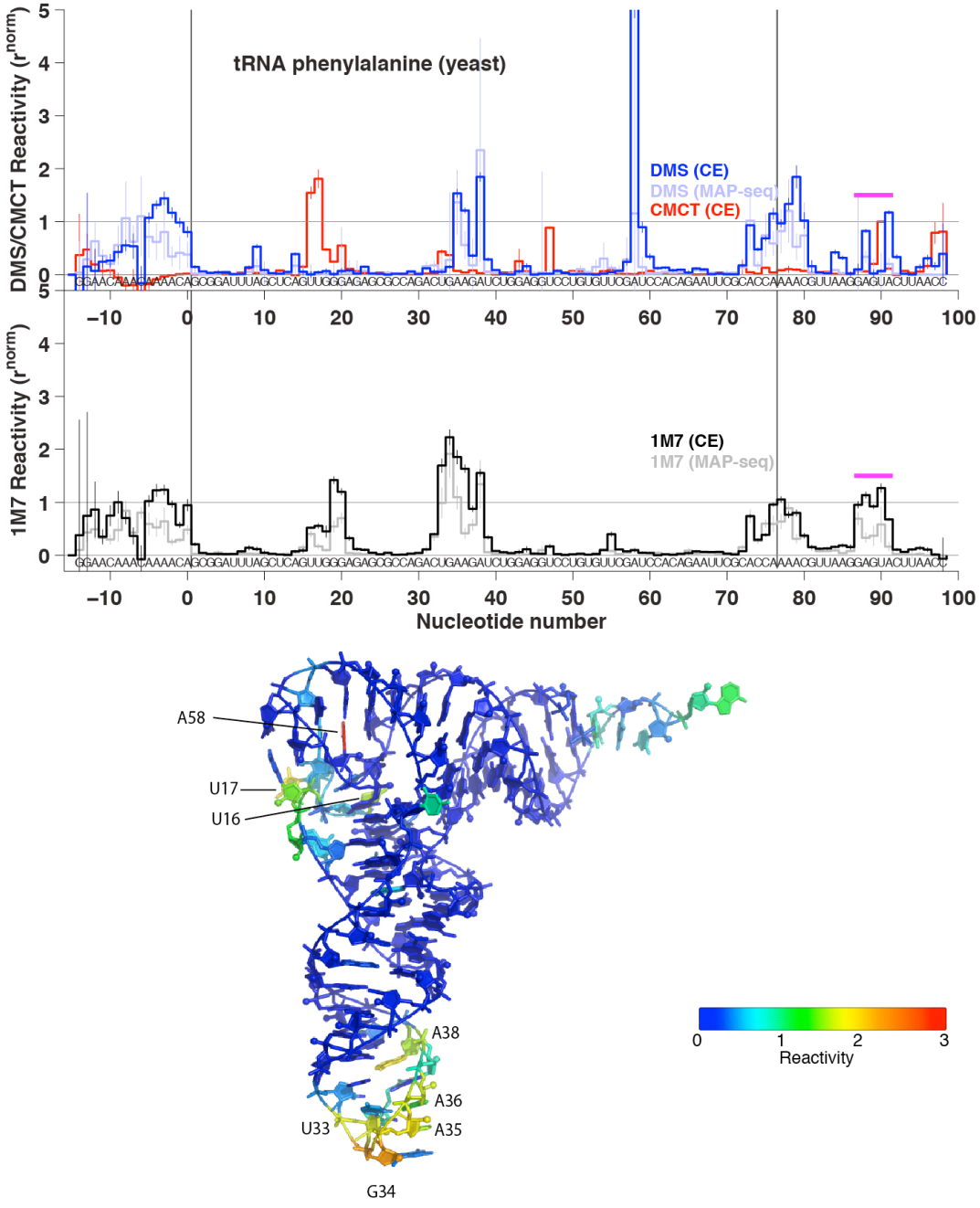
**SUPPORTING FIGURE 1. Example of data averaging and error estimation**. Output from `rdat_combine`function, available in HiTrace software. *Top panel*: heat-map representation of fourteen 1M7 measurements for the P4P6-2HP RNA enables rapid visualization of agreement and any sequence assignment errors; red-boxed residues are automatically determined outliers. *Bottom panel*: individual reactivity estimates (different colors) and averaged reactivity with error bars (black). Automatically determined outliers are marked with red circles.



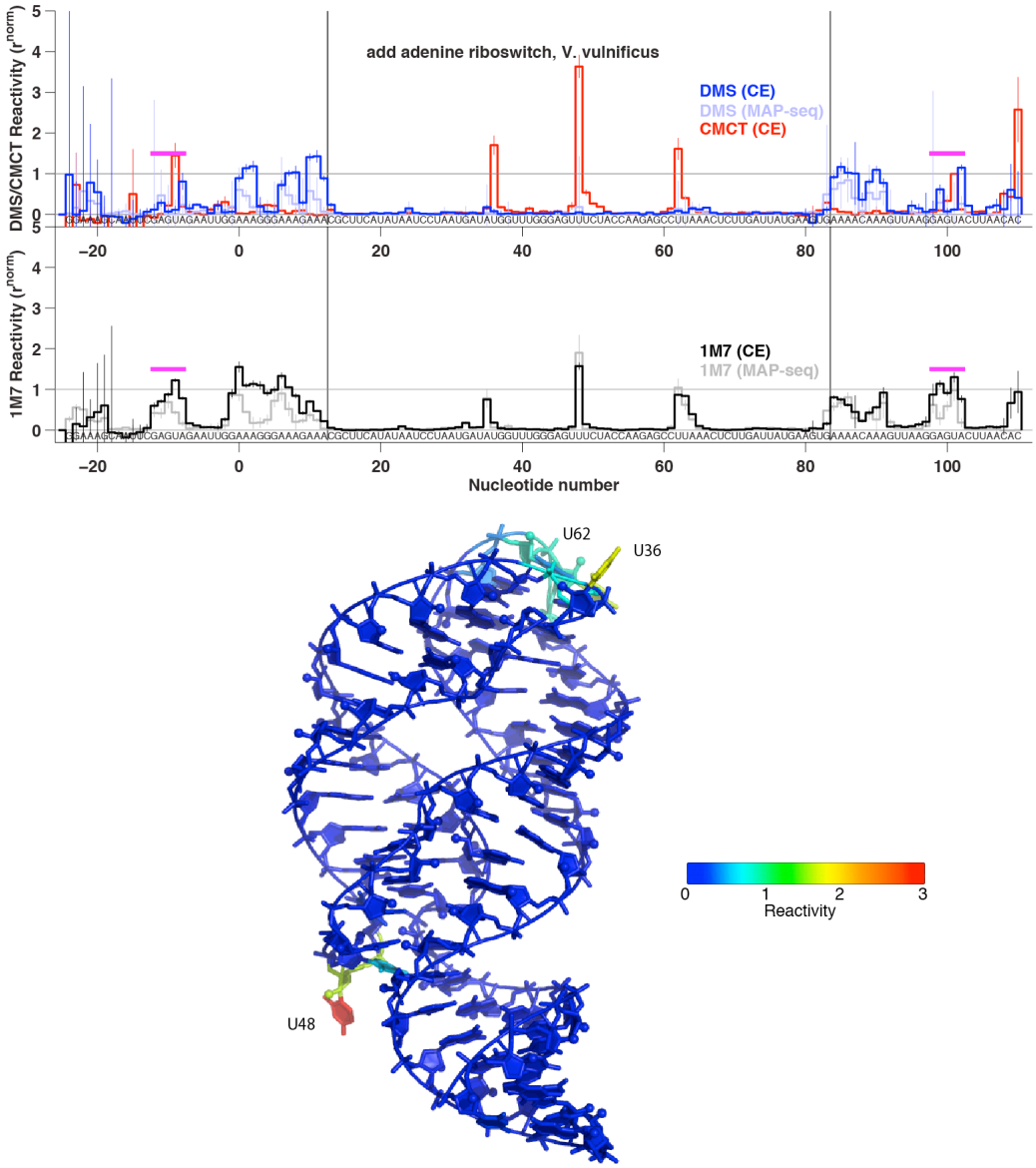
**SUPPORTING FIGURE 2. Proposed standardization brings data taken with varying chemical modifier concentrations into concordance**. Modifiers tested were DMS (A-B), 1M7 (C-D), and CMCT (E-F) on the P4-P6-2HP RNA. (A,C,E) HiTRACE peak fits to capillary electrophoresis traces give 'raw' peak areas with modified concentrations noted; black curves show background (no modifier) data. (B,D,F) Normalized reactivities after standardization (saturation correction, over-modification correction, background subtraction, and normalization to GAGUA reference hairpin at nucleotides 278-282).



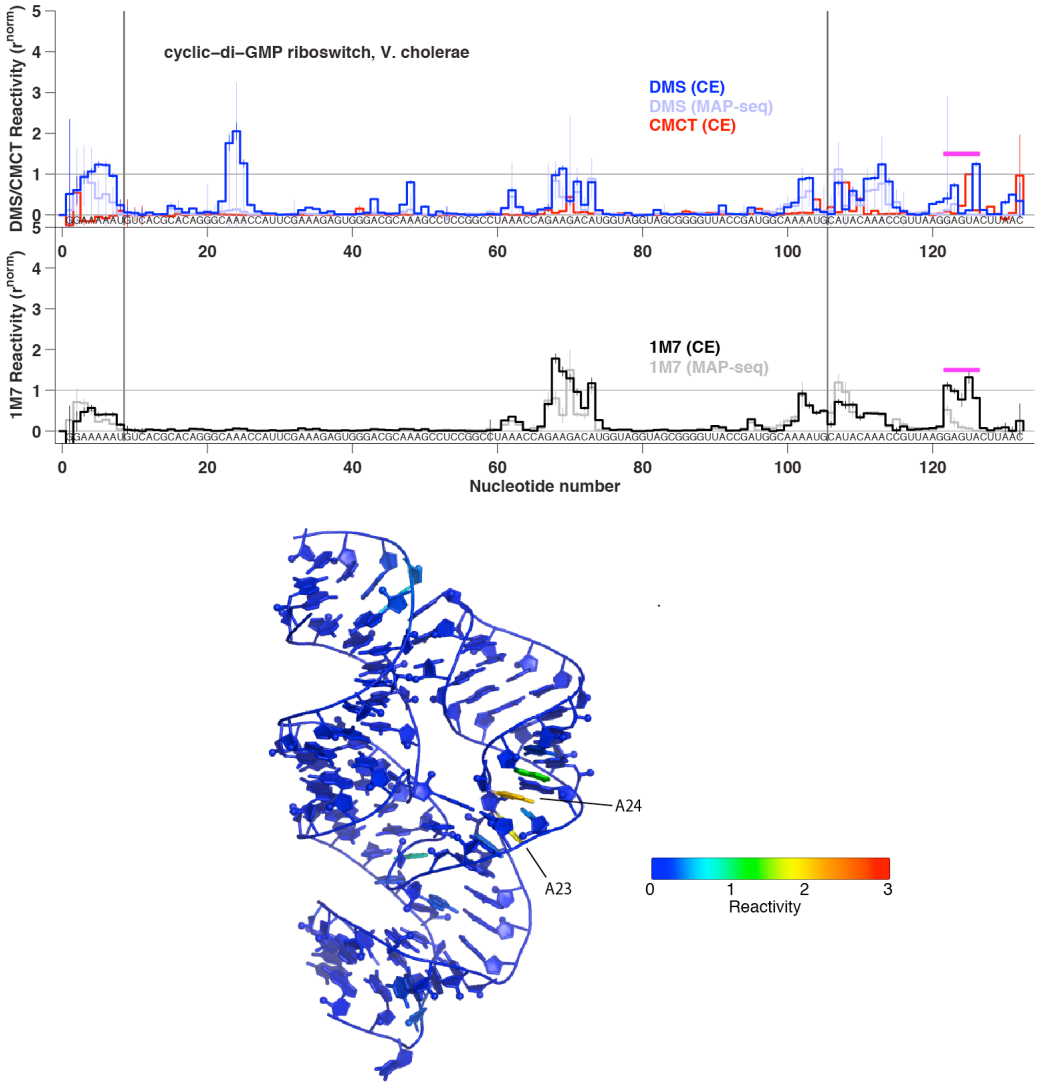
**SUPPORTING FIGURE 3. CE and MAP-seq data for unmodified tRNA(phe), *S. cerevesiae*.** In top panels, sequence of interest is between vertical black bars; GAGUA reference sequences marked in magenta.



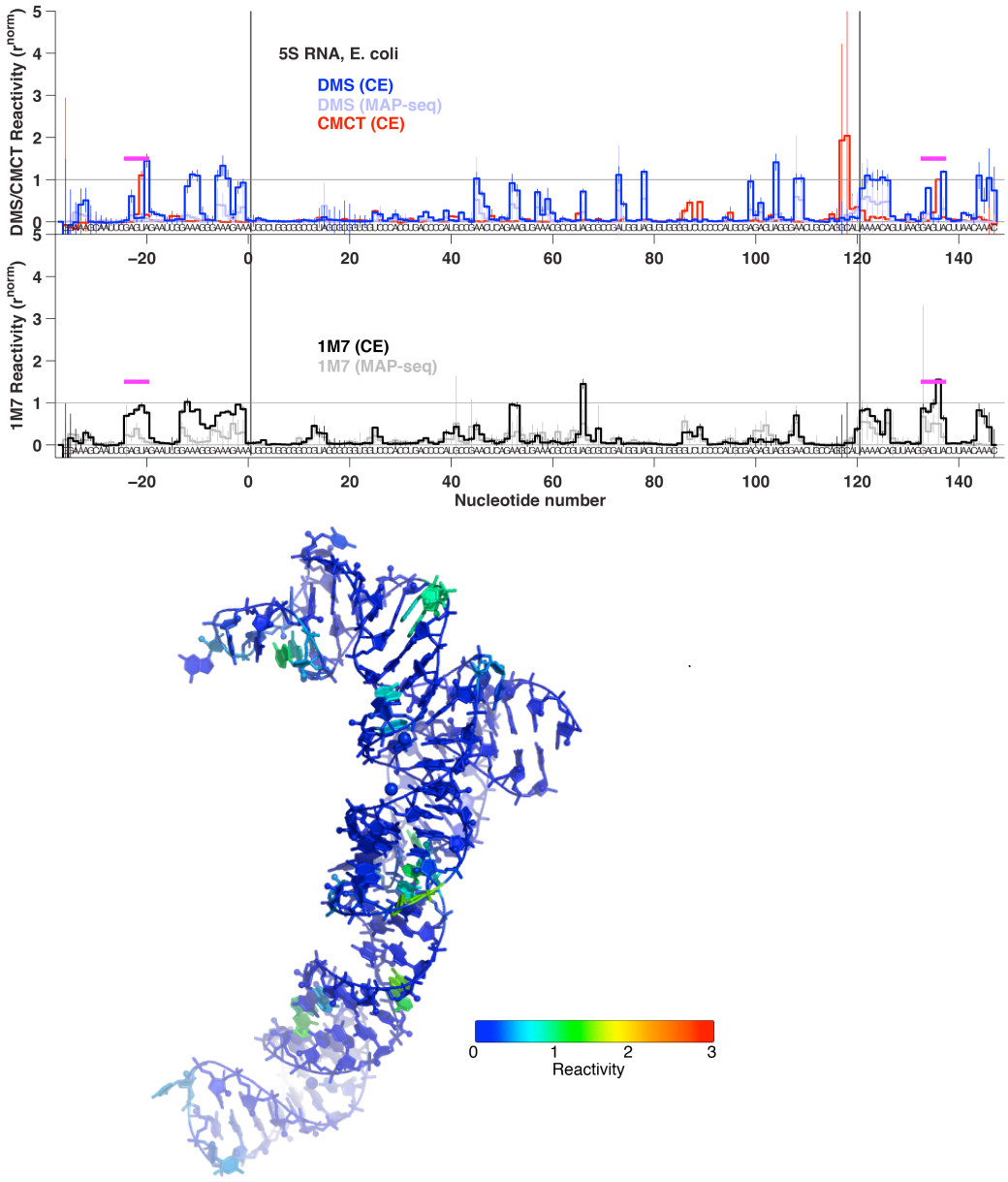
**SUPPORTING FIGURE 4. CE and MAP-seq data for ligand-binding domain of adenine riboswitch, *V. vulnificus.*** In top panels, sequence of interest is between vertical black bars; GAGUA reference sequences marked in magenta.



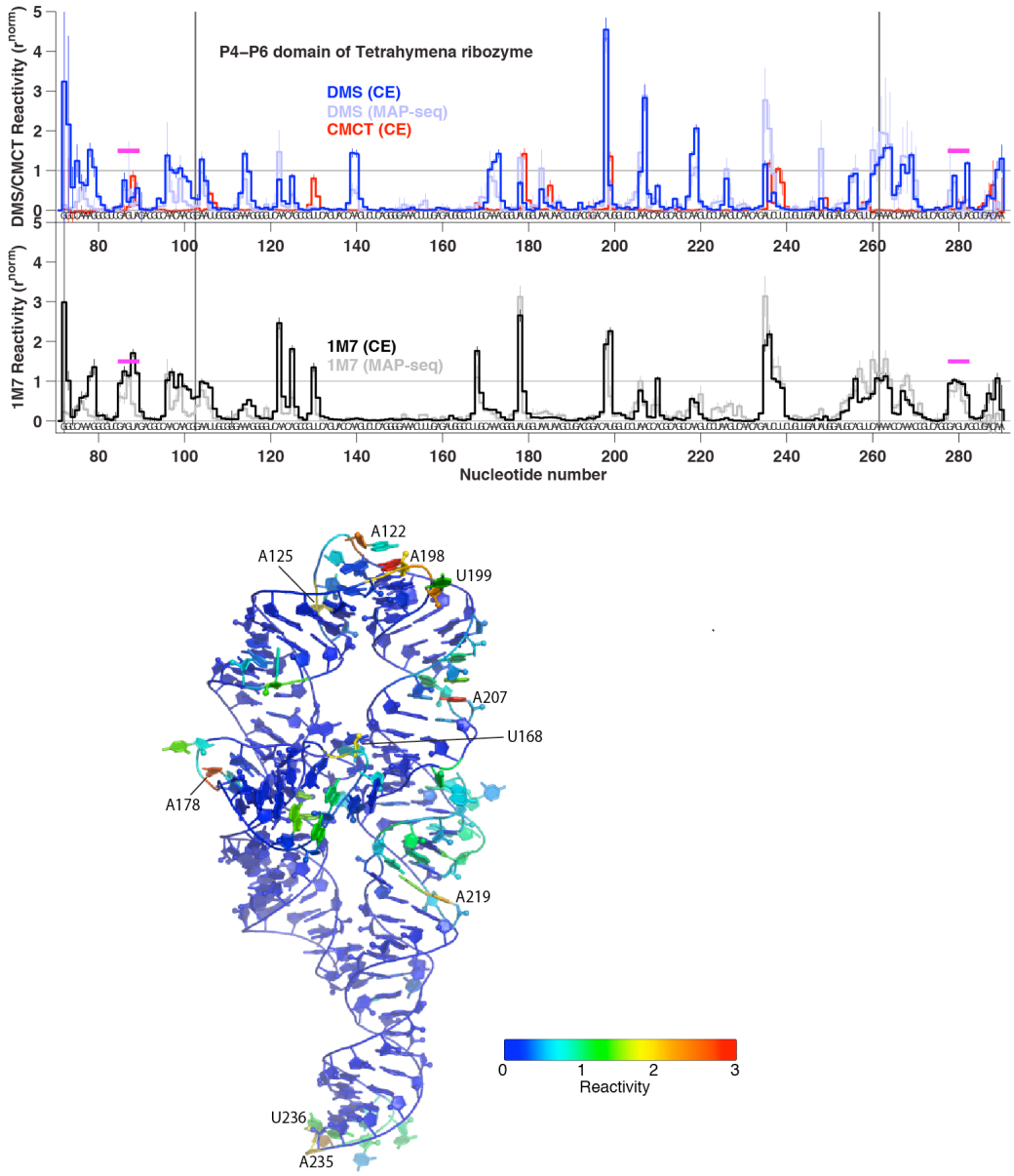
**SUPPORTING FIGURE 5. CE and MAP-seq data for ligand-binding domain of cyclic-di-GMP riboswitch, *V. cholerae.*** In top panels, sequence of interest is between vertical black bars; GAGUA reference sequences marked in magenta.



**SUPPORTING FIGURE 6. CE and MAP-seq data for 5S ribosomal RNA, *E. coli.*** In top panels, sequence of interest is between vertical black bars; GAGUA reference sequences marked in magenta.



**SUPPORTING FIGURE 7. CE and MAP-seq data for P4-P6 domain of the *Tetrahymena* ribozyme.** In top panels, sequence of interest is between vertical black bars; GAGUA reference sequences marked in magenta.



**SUPPORTING FIGURE 8. CE and MAP-seq data for ligand-binding domains of the glycine riboswitch, *F. nucleatum*.** In top panels, sequence of interest is between vertical black bars; GAGUA reference sequences marked in magenta.