Methods and Supplementary Information for

MOHCA-seq: RNA 3D models from single multiplexed proximity-mapping experiments

Clarence Yu Cheng[1], Fang-Chieh Chou[1], Wipapat Kladwang[1], Siqi Tian[1], Pablo Cordero[2], and Rhiju Das[1, 2, 3, *]

[1]Department of Biochemistry, [2]Biomedical Informatics Program, [3]Department of Physics, Stanford University, Stanford, California 94305, United States

*Correspondence should be addressed to R.D. (rhiju@stanford.edu).

This document contains the following sections:

Methods

RNA preparation

Double-stranded DNA templates for the RNAs of interest were constructed using PCR assembly with primers purchased from IDT (Integrated DNA Technologies)[4]. RNAs were transcribed at 37 °C for 3 hours in 320 μL reactions containing 32 pmol of dsDNA template, 100 mM Tris-HCl, pH 8.1, 200 mM $MgCl_2$, 3.5 mM spermidine, 0.1% Triton X-100, 40 mM DTT, 4% PEG 8000, 20 U T7 RNA Polymerase (New England Biolabs), 1 mM NTPs, and 0.5 mM 2´-NH_2-2´-deoxy-ATP (TriLink BioTechnologies). After purifying transcribed RNAs using RNA Clean-and-Concentrate columns (Zymo Research), RNAs were 5´-end labeled using a 5´-End-Tag kit and fluorescein maleimide (Vector Labs), then purified again using RNA Clean-and-Concentrate columns. The hydroxyl radical source, isothiocyanobenzyl-EDTA chelating Fe(III) (ITCB-Fe(III)•EDTA) (Dojindo Molecular Technologies, Inc.), was covalently attached to the 2´-NH_2 groups on the RNA backbone using a two-step process. First, to couple ITCB-EDTA to the RNA, the RNA was incubated with 0.5 mg ITCB-EDTA in 0.4 M KPO_4, pH 8.5 for 37 °C for 12-16 hours. Then, the coupling reactions were incubated with 67 mM $FeCl_3$ at room temperature for 15 minutes, after which 75-100 mM Na-EDTA, pH 8.0 was added to chelate excess Fe(III). After purifying with RNA Clean-and-Concentrate columns to remove excess reagents, the RNA was PAGE-purified using denaturing 8% polyacrylamide gels. Bands were located by scanning with a Typhoon imager (GE) for fluorescein fluorescence, and excised gel slices were immersed in RNase-free water in

nonstick tubes overnight at 4 °C to elute the RNA. RNAs were purified from the eluate using RNA Clean-and-Concentrate columns and stored at –20 °C.

Fragmentation and library preparation

Activation of radical source

Before activating the radical source to produce spatially localized hydroxyl radicals, 3-4 pmol amounts of folded RNA were prepared in 50 mM Na-HEPES, pH 8.0, and 10 mM $MgCl_2$, as follows. First the RNA was heated to 65 °C in HEPES buffer for 3 min, then cooled to room temperature for 10 min; then $MgCl_2$ was added and the RNA was heated to 50 °C for 5 min, then cooled to room temperature for 10 min. [For P4-P6, HEPES buffer and $MgCl_2$ were added concurrently, and the RNA was incubated for 10 min at room temperature. For ligand-bound glycine riboswitch samples, these incubation steps included 10 mM glycine. For ligand-bound adenosylcobalamin (AdoCbl) riboswitch samples, these incubation steps included 140 μM AdoCbl, and all steps until post-fragmentation ethanol precipitation (below) were performed in the dark.] After folding, the radical source was activated by adding a 100 mM sodium ascorbate stock to a final concentration of 10 mM. A control reaction was also prepared with deionized water added instead of ascorbate; this reaction was carried through all subsequent steps in parallel. After 5 to 30 min of incubation at room temperature (10 min was standard), 100 mM thiourea was added to a final concentration of 9.1 mM. The RNA fragments were ethanol precipitated as follows: First, 1 μL GlycoBlue (Life Technologies), 1/10 volume of 3 M sodium acetate, pH 5.2, and 3 volumes of 100% ethanol chilled on dry

ice were added to the reaction. Samples were spun down immediately in a tabletop microcentrifuge at maximum speed for 20 min, washed with ice-cold 70% ethanol, then spun down again at max speed for 10 min. The supernatant was removed by pipetting and the pellets were allowed to dry.

Repair of 3´ ends

To remove 3´-phosphates left by hydroxyl radical strand scission events in the RNA backbone[19, 20], the purified RNA fragments were treated with T4 polynucleotide kinase (T4 PNK) in conditions that promoted 3´-phosphatase activity[21]. The end-repair reactions contained 50 mM Na-MES, pH 6.0, 10 mM $MgCl_2$, 5 mM DTT, 5 μM ATP, and 10 units T4 PNK (NEB) and were incubated at 37 °C for 30 min. After end-repair, RNAs were ethanol precipitated as above, except with 0.5 μL GlycoBlue instead of 1 μL GlycoBlue.

Ligation of DNA tail

To prepare the fragmented RNA for reverse transcription, a preadenylated and 3´-blocked DNA tail (Universal miRNA cloning linker, NEB) was ligated to the 3´-end of the end-repaired fragments. Each ligation reaction contained 1x T4 RNA ligase buffer (NEB), 15% PEG 8000, 4–5 pmol DNA tail, and 200 U T4 RNA ligase 2 truncated, K227Q or KQ mutant (NEB). The reactions were incubated at 4 °C for 12 hours, followed by heat inactivation at 65 °C for 20 min. Ligated RNAs were purified from the

reaction using RNA Clean-and-Concentrate columns, which also removed excess DNA tail.

Reverse transcription with sequencing primers

After ligation of the DNA tail, RNAs were reverse transcribed with sequencing primers containing, in 5´ to 3´ order, a 5´-fluorescein modification, the Illumina TruSeq Universal adapter, 12-nucleotide barcodes (sequence-balanced in sets of 4 primers), and a primer for the DNA tail sequence on the 3´-end (Supplementary Table 1). Three of the four primers were used for reverse transcription of the ascorbate-treated sample and the remaining primer was used for reverse transcription of the no-ascorbate control sample. The 15 μL reverse transcription reactions contained 1x First Strand buffer (Life Technologies), 5 mM DTT, 0.8 mM dNTPs, and 120 U SuperScript III (Life Technologies) and were incubated at 55°C for 30 min. To degrade the RNA templates, 5 μL of 0.4 M NaOH was added and the samples were incubated at 90 °C for 3 min. After cooling on ice for 3 min, the cDNA samples were neutralized by addition of 1 μL of an acid quench (2 mL 5 M NaCl, 2 mL 2 M HCl, and 3 mL 3 M Na-acetate) and then purified by incubating with DynaBeads magnetic beads (Life Technologies) conjugated to double-biotin-labeled ssDNA complementary to the TruSeq adapter.

Ligation of second sequencing adapter

The second sequencing adapter (Supplementary Table 1), derived from the TruSeq Indexed adapter with a 5´-phosphate (to enable ligation to the cDNA) and a 3´-

phosphate (to block circularization) was ligated onto the 3´-ends of the cDNAs while they were still annealed to DNA-coated DynaBeads. The four cDNA samples for each RNA were pooled prior to the ligation reaction. The 50 μL ligation reactions contained 1.25 μM adapter, 1x CircLigase buffer (Epicentre), 50 μM ATP, 2.5 mM $MnCl_2$, 4% PEG 1500, and 250 U CircLigase I (Epicentre) and were incubated at 68 °C for 2 hours, followed by heat inactivation at 80 °C for 10 min. The samples were purified by magnetic separation, and a fraction of the sample was run on a capillary electrophoresis machine (Applied Biosystems) with co-loaded fluorescein-labeled standards and analyzed using HiTRACE[8] to estimate the concentration of ligated cDNA.

Sequencing

Sequencing using Illumina MiSeq or HiSeq instruments

The MOHCA-seq cDNA libraries were sequenced using 50-cycle MiSeq v2 kits on Illumina MiSeq instruments or using an Illumina HiSeq 2500 instrument (Elim Biopharmaceuticals). Beads harboring 25 fmol of single-stranded library fragments were mixed with 2.5 fmol of PhiX dsDNA control (Illumina) and EB buffer (Qiagen) to 5 μL total. Then 5 μL of 0.2 N NaOH was added and the fragments were eluted for 10 min at room temperature. The supernatant of the magnetic beads was diluted into chilled HT1 buffer (10 μL added to 990 μL HT1), and then diluted again (375 μL added to 225 μL HT1) before loading all 600 μL onto MiSeq kits following manufacturer instructions. Paired-end sequencing involved 51 and 25 sequencing cycles for the first and second reads, respectively. The resulting FASTQ data were processed with the MAPseeker

v1.2 software (https://github.com/DasLab/map_seeker), giving final text files in RDAT format corresponding to the number of fragments observed for each pair of possible cleavage and oxidation position (corresponding to 3´ and 5´ positions of cDNA fragments). These raw data have been deposited in the RNA Mapping Database (http://rmdb.stanford.edu/), with the following accession IDs: P4-P6 (TRP4P6_MCA_0001-0004), glycine riboswitch with 10 mM glycine (GLYCFN_MCA_0002) and with 0 mM glycine (GLYCFN_MCA_0003), AdoCbl riboswitch with 140 μM AdoCbl (RNAPZ6_MCA_0002) and with 0 μM AdoCbl (RNAPZ6_MCA_0003), and class I ligase (CL1LIG_MCA_0001-0003).

MOHCA-seq Data Analysis

General MOHCA-seq analysis framework

Analysis of MOHCA-seq data requires modeling backgrounds, modulation from reverse transcription attenuation, and sources of error. Here, we outline a general analysis framework for MOHCA-seq experiments, with the following two sections describing two independent statistical procedures developed to reach numerical solutions, which gave consistent proximity maps. The resulting proximity maps have been deposited in the RNA Mapping Database with the following accession IDs: P4-P6 (TRP4P6_MCA_0000), glycine riboswitch with 10 mM glycine (GLYCFN_MCA_0000) and with 0 mM glycine (GLYCFN_MCA_0001), AdoCbl riboswitch with 140 μM AdoCbl (RNAPZ6_MCA_0000) and with 0 μM AdoCbl (RNAPZ6_MCA_0001), and class I ligase (CL1LIG_MCA_0000).

A MOHCA-seq product stemming from a radical cleavage event at nucleotide *j* and a radical source at nucleotide *i* corresponds to a sequence *i+1* to *j–1* ligated between the two inserts necessary for paired-end Illumina sequencing. The frequency F_{ij} of such products is related to the proximity of *i* and *j* but is modulated by the actual distribution of radical sources, *e.g.*, primarily at adenosines for 2´-NH_2-dATP-incorporating transcripts. The frequency is also suppressed by signal attenuation for long sequence separations $j - i$ due to the possibility of reverse transcription termination between *i* and *j*. A master expression for these MOHCA-seq frequencies is:

$$F_{ij} = \sum_s p_i^s \left(1 - p_{i+1}^s\right)\left(1 - p_{i+2}^s\right)\ldots\left(1 - p_{j-1}^s\right) q_j^s \varepsilon(s) \qquad (1)$$

with *s* indexing the possible source positions (0, 1, … *N* with $s = 0$ corresponding to no source) and ε(*s*) the fraction of transcripts containing a source at *s*. Values p_i^s give the probability that a reverse-transcription-stopping event occurs at the nucleotide immediately 3´ to *i* for a transcript with source at *s*, and values q_j^s give the probability of a cleavage event between $j - 1$ and *j* for a transcript with source at *s*. This expression simplifies in the limit of no background processes and low radical cleavage and oxidation rates ($p_i^s, q_j^s << 1$, except at $p_i^{s=i} = 1$, corresponding to a reverse transcriptase stop at the site of radical source attachment). In that limit, eq. (1) reduces to:

$$F_{ij} = q_j^i \varepsilon(i) \qquad (2)$$

i.e. the observed frequencies provide a direct readout of the probability that a source at *i* leads to radical cleavage at position *j*. This was the limit assumed in prior gel-based MOHCA analysis based on reading out *i* through cleavage of phosphorothioate tags associated with the radical sources. In the present MOHCA-seq protocol, we found empirically that the reverse transcription readout of *i* led to a more complete portrait of the proximity map, including information at nucleotides *i* at which radical sources were not attached. Indeed, increasing the incorporation rate of radical sources or the time of ascorbate-induced radical damage produced higher signal-to-noise data sets with clear proximity map signals, despite bringing the analysis away from the regime of single-hit incorporation and damage (Supplementary Figs. 3 and 5). The extra information was derived from oxidative damage events that did not lead to backbone cleavage encapsulated in the term p_i^s (compare Fig. 1b to ref. 13; these non-cleaving events were previously invisible to gel-based analysis), but required a more advanced analysis to elicit the signal from the raw data.

In the general case, the number of observed frequencies F_{ij} is on the order of $N(N–1)/2$, whereas the total number of model parameters $\varepsilon(s)$, p_i^s , and q_j^s is substantially higher ($> 2\ N^2$), leading to an ill-posed problem. However, basic chemical considerations reduce the number of parameters. First, we assume that the cleavage fraction q_j^s is composed of a uniform 'background' rate of cleavage b_j that is independent of source *s*

(*e.g.*, due to inline attack during cleavage) and additional cleavage due to hydroxyl radical cleavage events (the desired MOHCA-seq signal), parameterized by π_j^s:

$$\text{Cleavage at } j \text{ from source } s = q_j^s = 1-(1-b_j)(1-\pi_j^s) \approx b_j + \pi_j^s \quad (3)$$

Second, we assume that the stopping events p_i^s are composed of a background rate r_i of oxidative damage (*e.g.*, due to solution radicals arising during ascorbate treatment) and then additional oxidation due to hydroxyl radical damage events:

$$\text{Stop at } i \text{ from source } s = p_i^s = 1-(1-r_i)(1-\rho_i\pi_i^s) \approx r_i + \rho_i\pi_i^s \quad (4)$$

In eqs. (3) and (4), a reduction in the number of parameters arises from assuming that oxidative damage events producing reverse transcription stops occur at rates proportional to backbone cleavage rates by a factor ρ_i. That is, π_i^s parameterizes the local effective concentration of radicals at *i* from source *s*, and the partitioning of these radicals into events that lead to cleavage (contributing to q_i^s) versus total damage that can terminate reverse transcription (contributing to p_i^s) is dependent on the chemical environment of the site *i* and not on source *s*. The total number of parameters thus reduces from greater than $2N^2$ to $N(N-1)/2$ for π_i^s and $3N$ for b_j, r_i, and $\varepsilon(s)$. By further enforcing positivity of each of these parameters and assuming that π_i^s is sparse, *i.e.* that each nucleotide gives non-negligible cleavage at a number of residues smaller than *N*, the number of parameters is reduced to well below the number of observables, and

the problem becomes well-posed. Requiring sparsity of the proximity map is similar to assumptions used in solvent flattening and other density modification approaches in crystallography[22].

Determining proximity information from raw observables still requires solving a complex system of non-linear equations. We found that direct least-squares optimization of the thousands of variables π_i^s to fit the observed F_{ij}, including Laplace priors to enforce sparsity, required hours even with state-of-the-art numerical optimizers. We instead developed two rapid, iterative strategies to carry out the solution (COHCOA and LAHTTE, described next) and used differences between the results to evaluate systematic errors in analysis assumptions.

Closure-based •OH COrrelation Analysis (COHCOA)

A Closure-based •OH COrrelation Analysis solves eq. (1) to determine a two-point correlation function that is directly read out by MOHCA-seq. In the limit that the fraction damaged at any single nucleotide is smaller than one,

$$\begin{aligned} F_{ij} &\approx \sum_s \left[p_i^s q_j^s - \sum_{i<m<j} p_i^s p_m^s q_j^s \right] \varepsilon(s) \\ &= \left\langle p_i q_j \right\rangle - \sum_{i<m<j} \left\langle p_i p_m q_j \right\rangle \end{aligned} \qquad (5)$$

where the bracket notation refers to a summation over sources:

$$\langle f \rangle = \sum_s \varepsilon(s) f^s \tag{6}$$

and terms beyond second order are neglected. The effects of the radical cleavage result in one-dimensional damage profile R_i and cleavage profile B_j and a two-point correlation function Q_{ij} which encodes the desired proximity map:

$$\begin{aligned} B_j &= b_j + \langle \pi_j \rangle \\ R_i &= r_i + \rho_i \langle \pi_i \rangle \\ Q_{ij} &= \rho_i \left[\langle \pi_i \pi_j \rangle - \langle \pi_i \rangle \langle \pi_j \rangle \right] \end{aligned} \tag{7}$$

leading to the equation

$$F_{ij} \approx R_i B_j + Q_{ij} - \sum_{i<m<j} \left(R_i R_m B_j + R_m Q_{ij} + R_i Q_{mj} + \rho_m Q_{im} B_j \right) \tag{8}$$

In the derivation of (8), we have dropped higher order terms corresponding to neglecting higher cumulants of the damage fuction (*e.g.*, the three-point cumulant $J_{imj} = \langle \pi_i \pi_m \pi_j \rangle - \langle \pi_i \pi_m \rangle \langle \pi_j \rangle - \langle \pi_i \pi_j \rangle \langle \pi_m \rangle - \langle \pi_j \pi_m \rangle \langle \pi_i \rangle + 2 \langle \pi_i \rangle \langle \pi_j \rangle \langle \pi_m \rangle$) to match the lowest order assumed in eq. (5). We also note that stops due to the possibility of more than one radical source attached to the transcript [neglected in the derivation above] can be modeled accurately, to first order, by including a rate ε(*i*) within the general 'background' stopping rate R_i (not shown). In general, all processes that lead to stops or cleavage across all transcripts are subsumed into R_i and B_j. As a corollary to this simplification,

however, this framework does not seek or enable deconvolution of the separate contributions to each of these background rates.

In practice, eq. (8) can give unphysical negative values if the subtracted summand becomes large. Following a strategy used in, *e.g.*, reference interaction site models for solvation, we 'close' the expansion by solving an equation system that is equal to (8) at lowest order, exact in certain limits, and guaranteed to give positive results:

$$F_{ij} = A_{ij}\left(R_i B_j + Q_{ij}\right)$$
$$A_{ij} = \prod_{i<m<j}\left(1 - R_m\right)\left(1 - \frac{R_i Q_{mj}}{F_{ij}}\right)\left(1 - \frac{\rho Q_{im} B_j}{F_{ij}}\right) \qquad \text{(9a,b)}$$

Intuitively, many of the observed products F_{ij} are due to uncorrelated cleavage events B_j and reverse transcription stop events R_i, which produce a 'plaid' background pattern. On top of this background is the desired two-point correlation signal Q_{ij} which is non-zero only when nucleotides *i* and *j* have both been chemically modified by the same proximal source. Modulating these signals is an attenuation factor A_{ij} which parameterizes the loss of signal, as a reverse transcriptase must polymerize from *j* back to *i*. This factor depends on the general background stop rate R_i, but also includes two additional terms representing the possibility for additional damage correlated with the observed cleavage event at *j* and the observed stopping event at *i*. (For simplicity and based on separate experiments, we fixed ρ, the ratio of chemical modification to backbone cleavage, as a constant at 2.5; changing this value from 1 to 5 gave indistinguishable results.) We note that the equation (9a,b) is exact in the case of negligible Q_{ij}; see references[23, 24].

The equations (9) are solved through iteration in a single script `cohcoa.m` available in the MAPseeker package. A starting estimate of R_i and attenuation matrix A_{ij} comes from data corresponding to cleavages in the 3´ flanking region, which is initially assumed to not give specific contacts with the target RNA domain. One-dimensional 'background' profiles R_i and B_j are determined which best fit the attenuation-corrected data F_{ij}/A_{ij}. Subtracting the resulting background matrix R_iB_j from observed F_{ij} results in an initial solution for Q_{ij}. Any point of Q_{ij} that is negative is reset to zero, and these R_i, B_j, and Q_{ij} give an updated solution for the attenuation matrix A_{ij}. New estimates of R_i and B_j are derived from fitting $F_{ij}/A_{ij} - Q_{ij}$ and the process is iterated until convergence. This procedure does not require assuming symmetry of the two-point correlation function Q_{ij}, but in the end returns an estimate of this matrix only for ($i < j$), where there are data F_{ij}. Propagating Poisson counting errors on F_{ij} in eq. 9b gives standard errors on A_{ij}, and combining these errors in quadrature with the errors on F_{ij} in eq. 9a gives final error estimates for the two-point correlation function Q_{ij}. Empirically, 20 or fewer iteration cycles lead to convergence for all data sets tested; 40 iterations have been used in this study to ensure convergence of final Q_{ij} values within 1% (taking less than 1 minute on a MacBook Pro 2.8 GHz Intel Core i7 running MATLAB 2012B). A comparison of raw counts and COHCOA-analyzed data is shown in Supplementary Figure 1a-b.

To visualize the data, we used a filter to remove points with a signal-to-noise ratio < 1 and applied a 2D smoothing algorithm (Supplementary Figure 1c). The analyzed

proximity maps presented in the figures were primarily generated by COHCOA analysis, with the exception of the comparison to LAHTTE analysis (described below) in Supplementary Figure 1d.

Likelihood Analysis of Hydroxyl-probed TerTiary contact Estimation (LAHTTE)

An alternative approach to obtain an estimate of the "proximity map" underlying the MOHCA-seq data is to simplify eq. (1) by only inferring the fraction of sequence fragments that are the result of source-induced cleavage in their 3´end and of reverse transcription stoppage due to the source itself at their 5´-end. This simplification drops contributions to the contact-map from sequence fragments that were a result from reverse transcription stoppage occurring before the source position due to natural reverse transcription drop-off or strand damage and is therefore an underestimate of the true contact probabilities. In this framework, we can rewrite F_{ij} as follows:

$$F_{ij} = \varepsilon_i p_i^j (1-R_i)(1-B_i)\prod_{i'=i+1}^{j}(1-R_{i'})(1-B_{i'})(1-\varepsilon_{i'})(1-p_i^{i'}) +$$
$$R_i(1-B_i)(1-\varepsilon_i)(1-p_i^j)\prod_{i'=i+1}^{j}(1-R_{i'})(1-B_{i'})(1-\varepsilon_{i'})(1-p_i^{i'}) +$$
$$R_i B_{ji}(1-B_i)(1-\varepsilon_i)(1-p_i^j)\prod_{i'=i+1}^{j}(1-R_{i'})(1-B_{i'})(1-\varepsilon_{i'})(1-p_i^{i'})$$

The first term corresponds to the event of interest: when the source at *i* is the cause of reverse transcription stoppage. The second and third terms correspond to reverse transcription stoppage due to background variables: in one event the 3´-end is cleaved by a source other than ε_i and in the other the cleavage event is due to background

cleavage (B_i). The terms $\prod_{i'=i+1}^{j}(1-p_i^{i'})$ are the result of excluding the possibility of reading out cleavage events caused by sources that are downstream (3´), which cannot be detected due to the nature of the MOHCA-seq protocol.

Letting X_{ij} be the number of sequence fragments observed from position *i* to *j*, we can write the likelihood of the data as:

$$L(p,B,R)=\prod_{i,j}F_{ij}^{X_{ij}}$$

and the log-likelihood:

$$LL(p,B,R)=\sum_{i,j}X_{ij}\log(F_{ij}) \tag{10}$$

We can then find a maximum likelihood estimator for *p* by differentiating (10) and setting it to zero. For brevity, we define the following variables:

$$\Gamma_{ij}=R_i(1-\varepsilon_i)(1-B_i)\left(1-\frac{B_j}{(1-B_j)}\right)$$

$$\Lambda_{ij}=\varepsilon_i(1-R_i)(1-B_i)-\Gamma_{ij}$$

Then, differentiating (10) with respect to *p* and setting to zero gives:

$$p_i^j = \max\left(0, \frac{X_{ij}\Lambda_{ij} - \left(\sum_{i'=1}^{i} X_{i'j} + \sum_{j'=j+1}^{N} X_{ij'}\right)\Gamma_{ij}}{2\Lambda_{ij}\left(\sum_{i'=1}^{i} X_{i'j} + \sum_{j'=j+1}^{N} X_{ij'}\right)}\right) \quad (11)$$

As in the general model for chemical mapping proposed previously[24], probabilities that are estimated to be negative are set to zero.

Notice that in (11), the term $\left(\sum_{i'=1}^{i} X_{i'j} + \sum_{j'=j+1}^{N} X_{ij'}\right)$ represents subtracting the background in the two dimensions indexed by *i* and *j*. Eliminating any of the summands reduces (11) to an equation similar to the solution for reactivity probabilities in chemical mapping. This "two-dimensional background" is readily observed in the raw data and is the primary structure exploited by the COHCOA analysis described above.

To calculate the variance of our maximum likelihood estimator, we use the second derivative of the log-likelihood with respect to *p*. Letting $C_{ij} = p_i^j \varepsilon_i (1 - R_i)(1 - B_i) - (1 - p_i^j)\Gamma_{ij}$, we then have that:

$$Var(p_i^j) = \frac{1}{\frac{\left(\sum_{i'=1}^{i} X_{i'j} + \sum_{j'=j+1}^{N} X_{ij'}\right)}{(1 - p_i^j)^2} - \left(\frac{\Lambda_{ij}}{C_{ij}}\right)^2}$$

All that remains is to obtain estimators for ε, R, and B. However, due to the appearance of ε_i, R_i, and B_i in multiple F_{ij} terms in (10), the likelihood function may have at most N different solutions for each ε_i, R_i, and B_i, where N is the sequence length. It is therefore desirable to obtain the background variables using additional data. For example, ε can be estimated by comparing the MOHCA-seq data to a no-ascorbate control using the standard formalism for obtaining reactivity probabilities in chemical mapping experiments. Furthermore, a solution hydroxyl radical cleavage assay can help estimate R and B. As a final note, we found, surprisingly, that the maximum likelihood estimator of p does not depend strongly on the values of background variables ε, R, and B: it seems that the estimator is sensitive only to specific variable orderings (*e.g.* $\varepsilon > R > B$, $\varepsilon > R < B$, etc.). An example of LAHTTE performed on a P4-P6 dataset is shown in Supplementary Figure 1d.

Quantile normalization of RNAs probed in multiple states

For RNAs probed in multiple states, such as the glycine and AdoCbl riboswitches, we found that differences in the proximity maps were best visualized using difference maps. To produce these maps, we first quantile-normalized COHCOA-analyzed MOHCA-seq datasets corresponding to the two states of interest. Quantile-normalized maps are shown in Figure 2a-b and d-e as well as in Supplementary Figures 11a-b and 12a-b. We then subtracted the ligand-free data from the ligand-bound data. Lastly, we used a filter to remove points with a signal-to-noise ratio < 1 and applied a 2D smoothing algorithm.

Supplementary Figures 11c and 12c show the resulting difference data plotted on heatmaps with positive and negative signal cutoffs of three standard deviations.

Mutate-and-map

High-confidence secondary structures derived from mutate-and map (M^2) experiments were used as inputs to computational modeling. We previously collected M^2 data on the AdoCbl riboswitch in the presence of 60 μM AdoCbl [RMDB ID RNAPZ6_1M7_0002] and separately performed M^2 on the class I ligase. The data were analyzed as described previously[4, 25]. Supplementary Figures 7 and 9 present M^2 data and analysis for the class I ligase and AdoCbl riboswitch, respectively.

Computational modeling

***De novo* RNA modeling with MOHCA-seq constraints**

We used the Rosetta software (version r56277) to model all RNAs of interest in three steps. Modeling for the AdoCbl riboswitch was performed with some differences, as noted below. First, based on the secondary structures of the RNAs from crystal structures or chemical mapping experiments, we pre-assembled the helix regions of the RNAs using fragment assembly of RNA with full atom refinement (FARFAR), using the python script `helix_preassemble_setup.py`. For each helical region, 100 FARFAR models were generated, with the following sample Rosetta command line:

```
rna_denovo -nstruct 100 -params_file helix.params -fasta helix.fasta
-out:file:silent helix0.out -include_neighbor_base_stacks  -
minimize_rna true -rna::corrected_geo -score:rna_torsion_potential
```

```
RNA11_based_new -geom_sol_correct_acceptor_base -
chemical::enlarge_H_lj -score:weights rna/rna_helix -cycles 1000 -
output_res_num 136-142 221-227 -output_res_num  136-142 221-227
```

Second, we performed low-resolution modeling using the pre-assembled helices and tertiary constraints derived from maxima in the 2D MOHCA-seq proximity map. Tertiary constraints were determined by selecting peaks from early analyzed data that were (1) distinguishable from the local background signal by unbiased inspection and (2) not attributable to secondary structure. This produced a list of pairs of residues that were suggested to be spatially proximal by the MOHCA-seq data (tabulated in Supplementary Table 2). We sorted the selected pairs for each RNA into strong and weak hits based on the apparent intensity of the signal relative to the background. For each pair of residues showing a strong MOHCA-seq hit, we constrained the distance between the O2´ of first residue and C4´ of the second residue using a potential of the following form (see graph in Supplementary Fig. 6):

$$S(x) = 3x^2 - 2x^3$$

$$E(d) = \begin{cases} 4(1 - S(d/15)) & \text{if } 0 \le d < 15 \\ 4S(d/15 - 1) & \text{if } 15 \le d < 30 \\ 4 + 36S(d/30 - 1) & \text{if } 30 \le d < 60 \\ 40 & \text{if } d \ge 60 \end{cases}$$

Here S is the smoothstep function, d is the distance between the atom pair, and E is the constraint potential in Rosetta unit. For residue pairs with a weak MOHCA-seq signal, we applied a constraint potential of the same shape but weaker amplitude (1/5 of the original potential). With the constraints and the pre-assembled helices, we performed a

fragment assembly of RNA (FARNA) simulation to generate a large number of low-resolution Rosetta models (ranging from 10,000 to 60,000). For cases in which other experimental constraints (hydroxyl radical footprinting for the P4-P6 domain and class I ligase; tertiary information from mutate-and-map on the P4-P6 domain) were available, these constraints were also included in the modeling. A sample Rosetta command line is shown below:

```
rna_denovo -nstruct 500 -params_file rna_params.params -fasta
rna_fasta.fasta  -out:file:silent rna.out -
include_neighbor_base_stacks  -minimize_rna false -native
rna_native.pdb  -in:file:silent helix0.out helix1.out helix2.out
helix3.out helix4.out helix5.out helix6.out -input_res  1-8 19-26 12-
17 122-127 42-46 84-88 55-61 113-119 62-66 72-76 79-82 89-94 111-112 -
staged_constraints -cycles 20000 -output_res_num  1-127
```

In the final step, we refined models from the initial run that had low Rosetta low-resolution scores (within the lowest 1/6 of the models) using the high-resolution Rosetta score to obtain final minimized models. This was achieved using the `rna_minimize` Rosetta application:

```
rna_minimize -native rna_native.pdb -cst_fa_file constraint -
params_file rna_params.params -skip_coord_constraints -in:file:silent
0.silent -out:file:silent rna_min.out
```

To find representative models, we clustered the lowest-energy models (0.5% of the total number of the unrefined models), with a threshold (based on all-heavy-atom RMSD) chosen so as to give 1/6 of the clustered models in the most populous cluster, as in

prior work[10]. The clustering was achieved by first finding the model (the 'cluster center') with the largest number of neighboring models within the RMSD threshold, then assigning these models to the first cluster (cluster 0). This process was then repeated to cluster all of the remaining models. Clustering was performed using Rosetta, with the following command line:

```
cluster -in:file:silent silent.out -in:file:fullatom -
out:file:silent_struct_type binary_rna -export_only_low false -
out:file:silent cluster.out -cluster:radius 7
```

The RMSD threshold described above was used as an estimate of the precision of the structure ensemble from the modeling. The accuracy of the ensemble was estimated using the median of the RMSD between each of the models in the most populous ensemble to the gold-standard crystal structure. The accuracy and precision of modeling for the AdoCbl riboswitch were determined slightly differently, as described below, to allow direct comparison to prior modeling efforts in the RNA-puzzle challenge. In addition, we calculated the percentage of strong and weak MOHCA-seq constraints satisfied by our models (averaged over all models in each cluster) and the crystal structure. We defined a constraint as 'satisfied' if the O2′ of the first residue was within 30 Å of the C4′ of the second residue. The MOHCA-seq-guided models satisfied 80% or more of the strong constraints and 60% or more of the weak constraints. As expected, the models generally satisfied a greater percentage of constraints than the crystal structures; however, the percent of constraints satisfied was not correlated with the

accuracy or precision of the models. The details of the modeling results and the statistics computed as discussed above are available in Supplementary Table 3.

For the AdoCbl riboswitch, we did not pre-assemble helices and instead used prebuilt RNA fragments that we previously generated during the RNA-puzzles challenge. These fragments consisted of three models each of the P1 through P6, P7 through P8, and P10 through P11 regions. We modeled the AdoCbl riboswitch with three distinct combinations of these fragments, each containing one fragment of each region. To estimate the precision of modeling for the AdoCbl riboswitch in the ligand-bound and ligand-free states, we calculated the average pairwise all-heavy-atom RMSD between the cluster centers of the largest clusters for the three modeling setups. The accuracy of the modeling for each ligand-binding state was determined by the median of the accuracies for the three modeling setups. The prior models generated for the RNA-puzzle 6 challenge can be viewed on the RNA-puzzles website (http://paradise-ibmc.u-strasbg.fr/rnapuzzles/index.html).

Supplementary Figure Legends

Supplementary Figure 1. MOHCA-seq data analysis.
(a) Raw counts for example P4-P6 MOHCA-seq data. Following paired-end sequencing, the MAPseeker software is used to align the reads to the sequence of the RNA that was probed. The counts are recorded in an RDAT file, which is analyzed by Closure-based •OH COrrelation Analysis (COHCOA). **(b)** COHCOA analysis after 40 iterations on example P4-P6 data. A full description of the COHCOA analysis can be found in the Methods section. **(c)** Final analyzed example P4-P6 proximity map. A filter is applied to remove points with signal-to-noise ratio < 1, and a 2D smoothing algorithm is applied. **(d)** LAHTTE analysis of the same example P4-P6 MOHCA-seq data. A full description of the LAHTTE analysis can be found in the Methods section.

Supplementary Figure 2. Full MOHCA-seq proximity maps, including 5´ and 3´ buffer regions.
(a-f) Complete MOHCA-seq proximity maps, including 5´ and 3´ buffer regions, which are excluded from the proximity maps shown in the main text. All RNAs except for P4-P6 include 5´ and 3´ reference hairpins used for normalizing chemical mapping data from other techniques (TH Mann, WK, RD, personal communication) (Supplementary Table 1).

Supplementary Figure 3. Variation of radical source incorporation rate.
MOHCA-seq data for P4-P6. Ratio of concentrations of 2´-NH_2-2´-dATP to ATP in transcription reaction: **(a)** 0; **(b)** 0.2; **(c)** 0.5 (standard); **(d)** 1.25. All fragmentation reactions were performed for 10 minutes.

Supplementary Figure 4. Alternative nucleotide attachment sites for the radical source.
MOHCA-seq data for P4-P6. Modified nucleotide triphosphate included in transcription reaction at molar ratio of 0.5 to standard NTP: **(a)** 2´-NH_2-2´-dATP; **(b)** 2´-NH_2-2´-dUTP; **(c)** 2´-NH_2-2´-dGTP; **(d)** 2´-NH_2-2´-dCTP. All fragmentation reactions were performed for 30 minutes. All data was collected in one Illumina MiSeq run using a 50-cycle MiSeq Reagent Kit v2.

Supplementary Figure 5. Variation of fragmentation reaction time.
MOHCA-seq data for P4-P6. Fragmentation reaction time: **(a)** 0 min (no ascorbate added); **(b)** 5 min; **(c)** 12.5 min; **(d)** 30 min. All fragmentations reactions used a 2´-NH_2-2´-dATP:ATP ratio of 0.5. The P4-P6 construct used for this experiment included flanking hairpins 5´ and 3´ reference hairpins used for standardizing chemical mapping data (TH Mann, WK, RD, personal communication) (Supplementary Table 1). All data was collected in one Illumina MiSeq run using a 50-cycle MiSeq Reagent Kit v2.

Supplementary Figure 6. Rosetta scoring potential used to incorporate MOHCA-seq constraints in modeling.
Shown is the scoring potential for strong MOHCA-seq hits, which was generated using the smoothstep function as described in the Methods. The scoring potential for weak MOHCA-seq hits was the same except with 5-fold lower amplitude.

Supplementary Figure 7. Mutate-and-map (M^2) analysis of class I ligase using 1M7 modifier.
(a) Mutate-and-map dataset for 1M7 modification across 120 single mutations along the class I ligase sequence. **(b)** Z-score contact map extracted from (a). **(c)** Secondary structure prediction and **(d)** bootstrap support matrix using M^2 data. In (b) and (d), the crystallographic secondary structure is overlaid as cyan and green circles, with an alternative P3 helix predicted by M^2 data overlaid as red circles. When SHAPEknots[26] was used to predict the pseudoknots in the full sequence, only the P2 pseudoknot was recovered. However, SHAPEknots successfully predicted the P3 helix when the P1 helix was omitted.

Supplementary Figure 8. Comparison of class I ligase crystal structure and knotted and unknotted cluster centers.
(a) Crystal structure of the class I ligase. **(b)** Cluster center of a knotted cluster of MOHCA-seq-guided models; the 3´-end of the RNA is knotted (black arrow). **(c)** Cluster center of an unknotted cluster of models.

Supplementary Figure 9. Mutate-and-map analysis of the adenosylcobalamin riboswitch, acquired during the sixth RNA-puzzles structure prediction trials.
(a) Mutate-and-map dataset for 1M7 modification across 168 single mutations along the AdoCbl riboswitch sequence in the presence of 60 μM AdoCbl ligand. Mutants showing poor data quality are marked by red bars. **(b)** Z-score contact map extracted from (a). **(c)** Secondary structure prediction and **(d)** bootstrap support matrix using M^2 data. In (b) and (d), the crystallographic secondary structure is overlaid as cyan circles.

Supplementary Figure 10. Comparison between MOHCA-seq-guided models of the adenosylcobalamin riboswitch using different initial RNA fragment sets.
(a) Crystal structure of *S. thermophilum* AdoCbl riboswitch (PDB ID 4GXY) and MOHCA-seq models generated using the three distinct sets of prebuilt RNA fragments (labeled setups 1-3) in the presence of **(b)** 140 μM AdoCbl ligand or **(c)** 0 μM AdoCbl ligand. Models shown include cluster center (opaque) and four other models from the largest cluster.

Supplementary Figure 11. Difference map comparison of the ligand-bound and ligand-free states of the glycine riboswitch.
MOHCA-seq data for glycine riboswitch with **(a)** 10 mM glycine and **(b)** 0 mM glycine. **(c)** Difference map between the ligand-bound and ligand-free MOHCA-seq data.

MOHCA-seq hits enhanced in the ligand-bound state are yellow (positive) and hits enhanced in the ligand-free state are cyan (negative).

Supplementary Figure 12. Difference map comparison of the ligand-bound and ligand-free states of the adenosylcobalamin riboswitch.
MOHCA-seq data for AdoCbl riboswitch with **(a)** 140 μM AdoCbl and **(b)** 0 μM AdoCbl. **(c)** Difference map between the ligand-bound and ligand-free MOHCA-seq data. MOHCA-seq hits enhanced in the ligand-bound state are yellow (positive) and hits enhanced in the ligand-free state are cyan (negative).

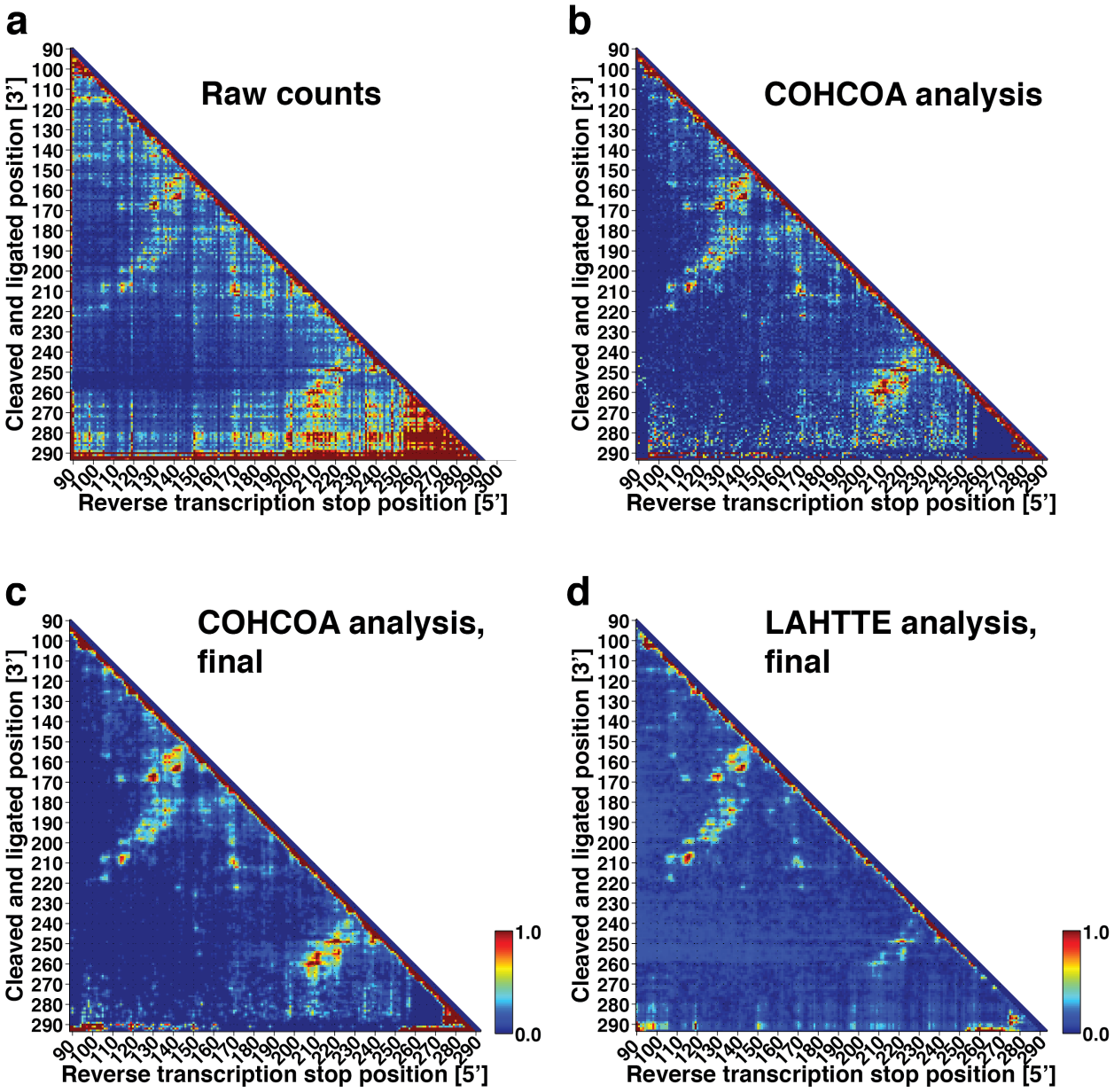
Supplementary Table 1. Sequences of RNAs, single-stranded DNA ligation adapters, and sequencing primers.

Supplementary Table 2. Pairwise MOHCA-seq constraints used for *de novo* modeling.

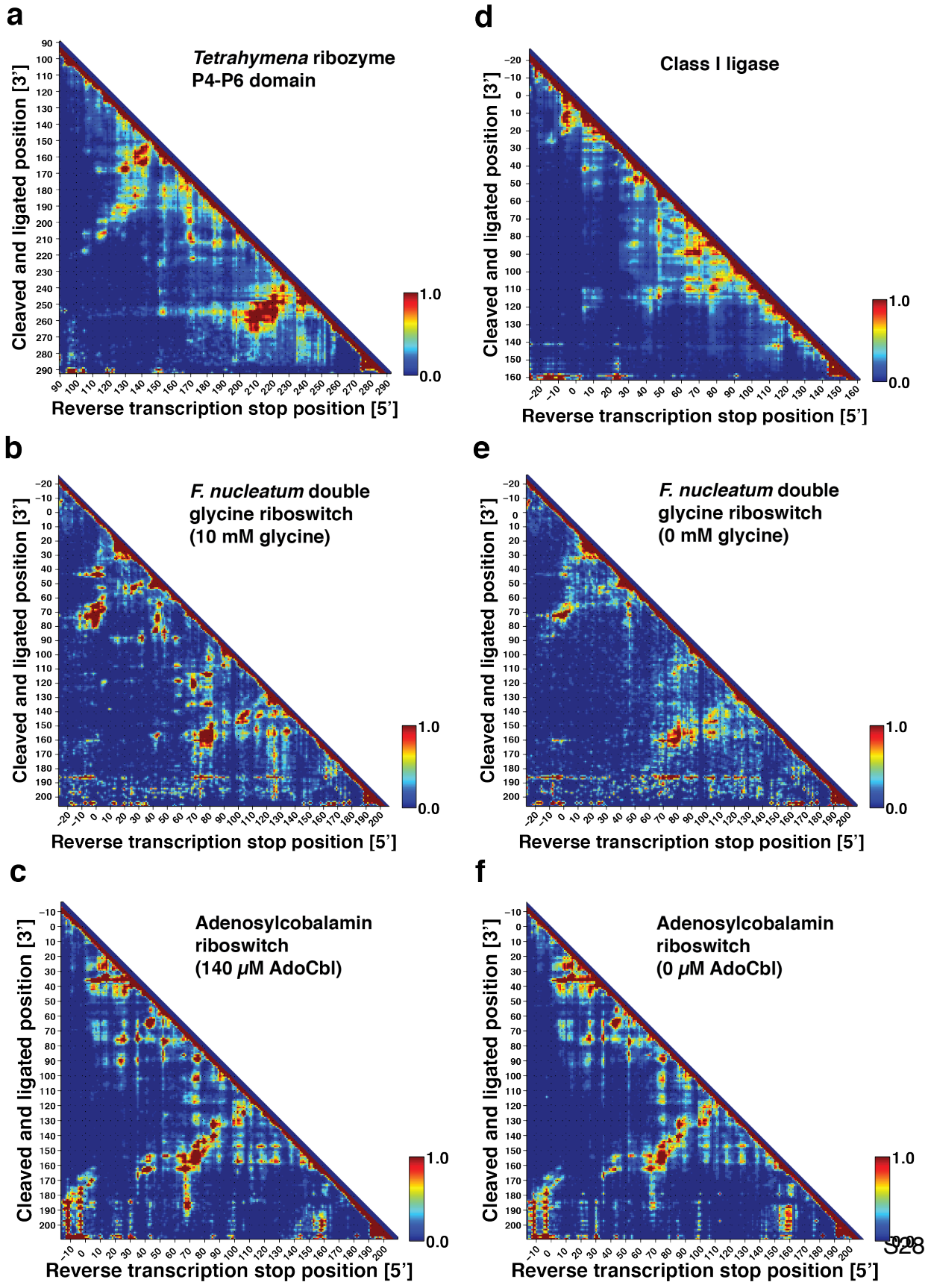
Supplementary Table 3. Results of *de novo* modeling incorporating MOHCA-seq constraints.

Supplementary Figures 1–12

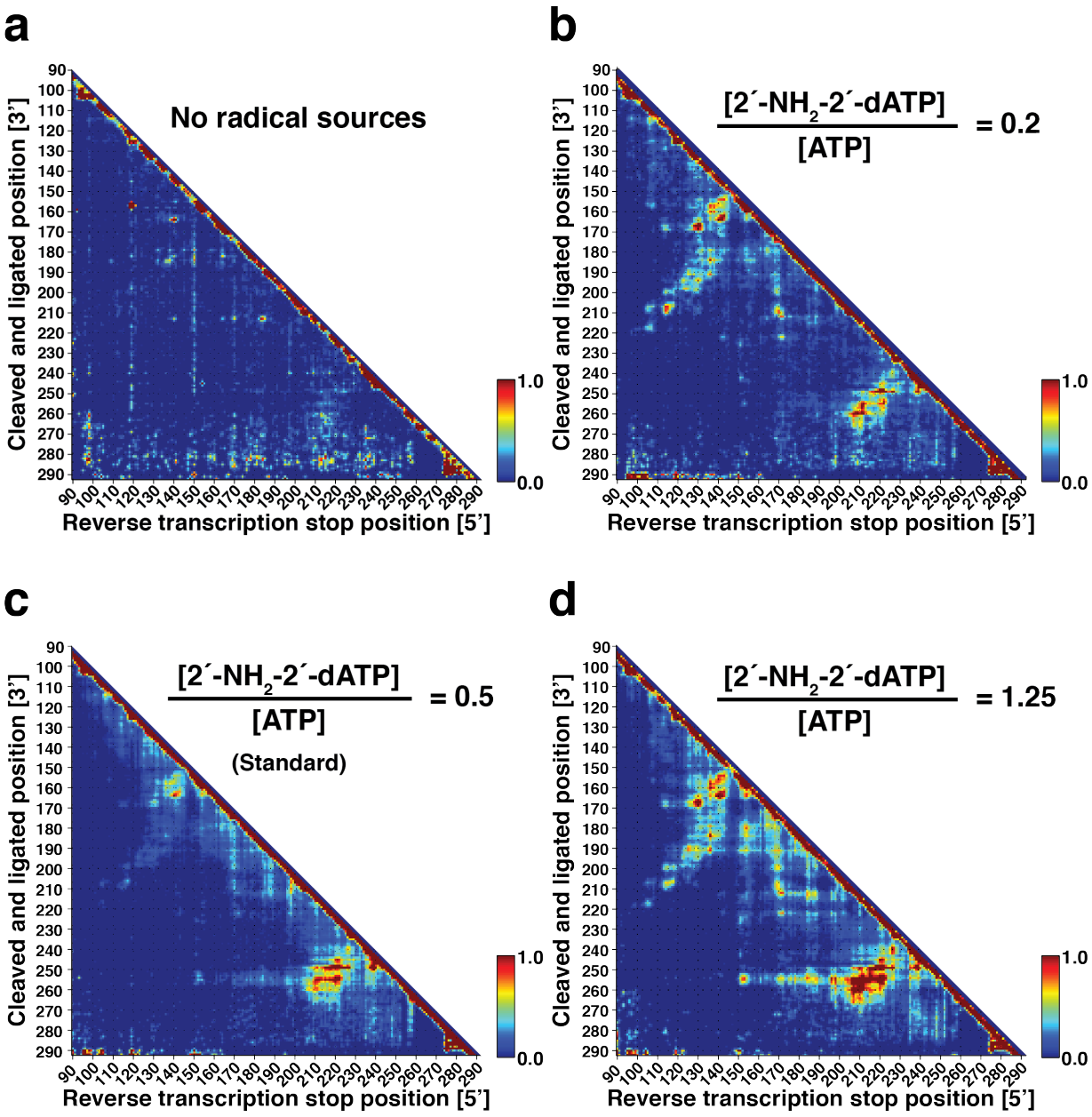
Supplementary Figure 1. MOHCA-seq data analysis



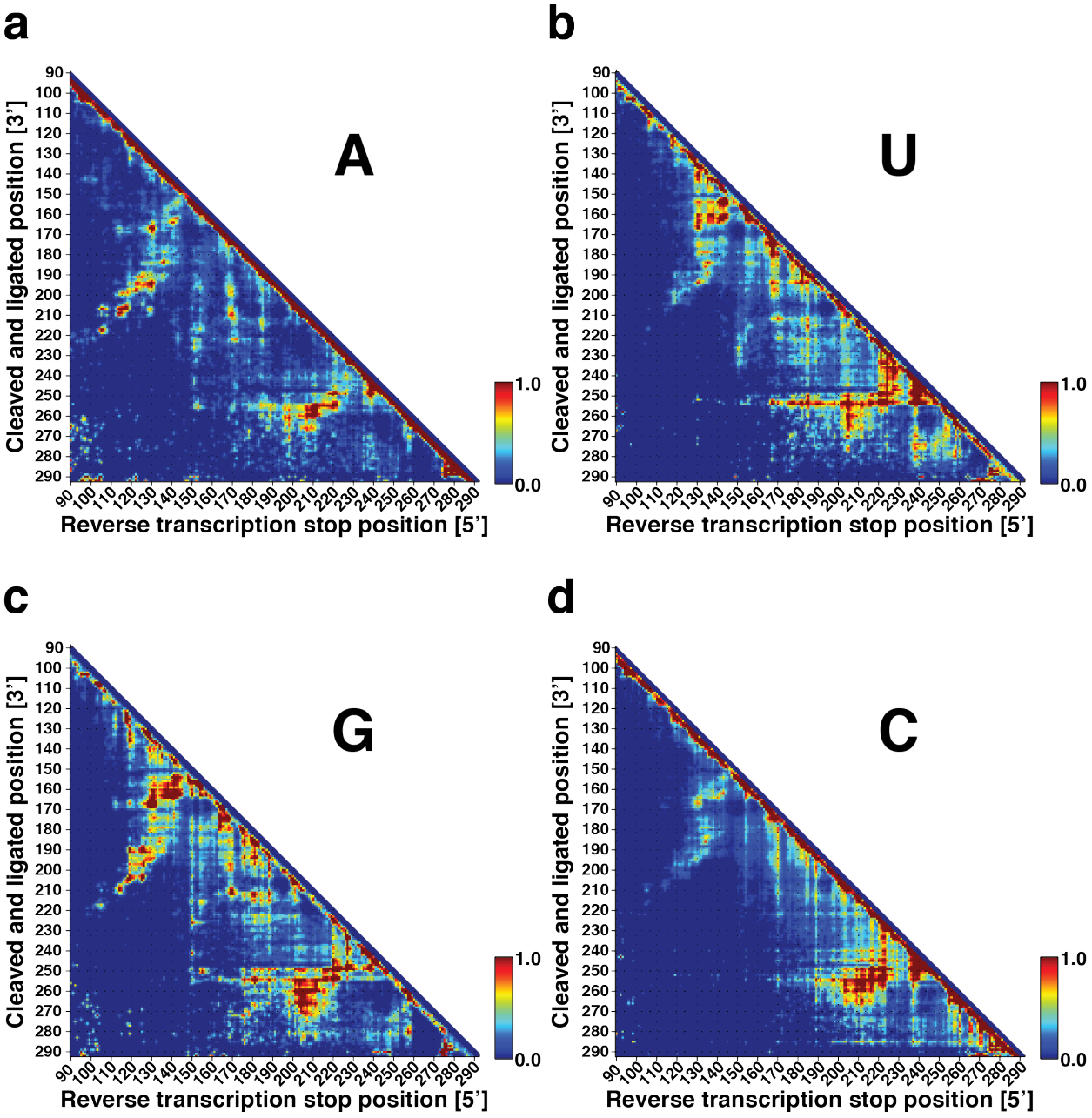
Supplementary Figure 2. Full MOHCA-seq proximity maps, including 5´ and 3´ buffer regions



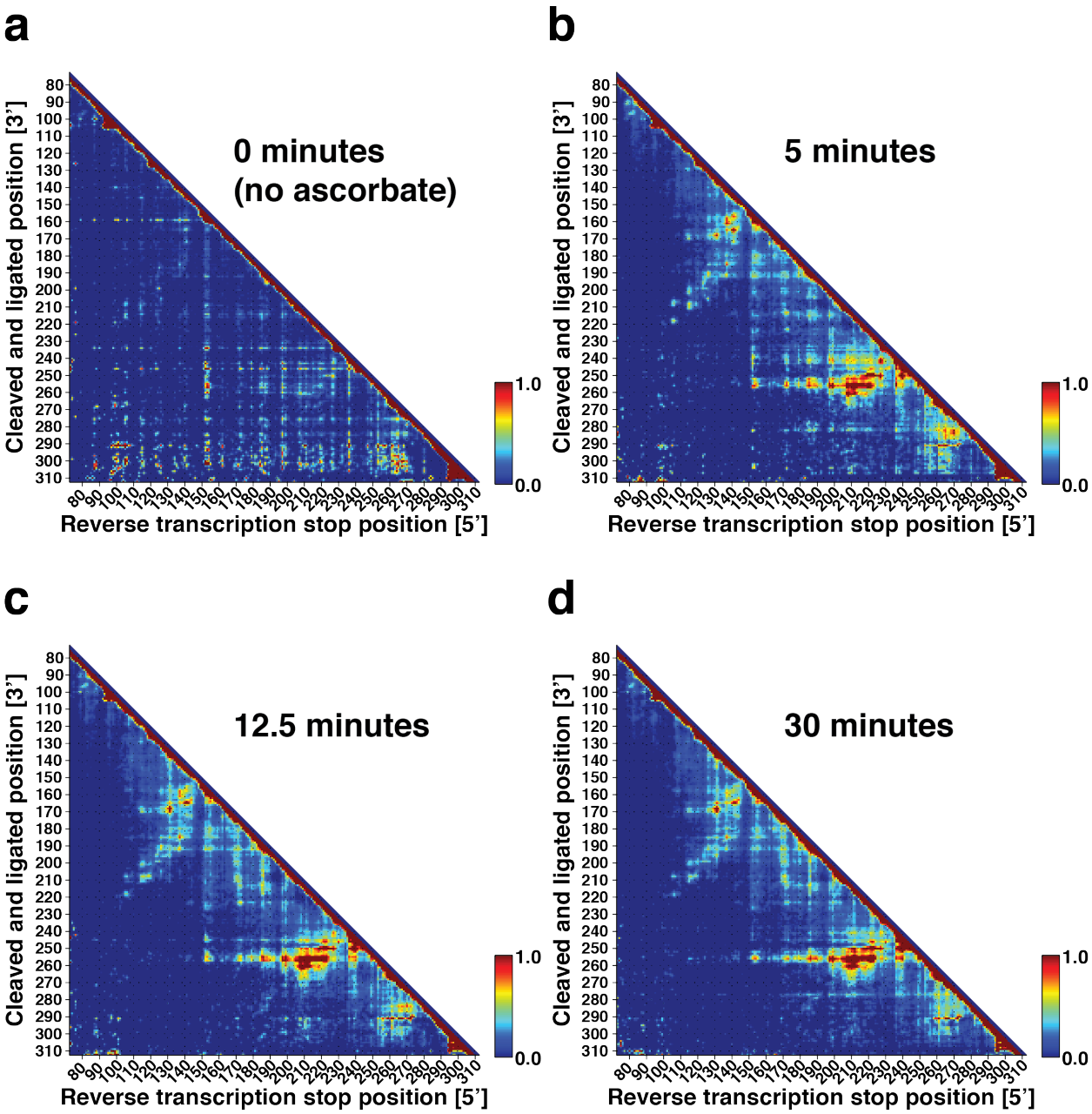
Supplementary Figure 3. Variation of radical source incorporation rate



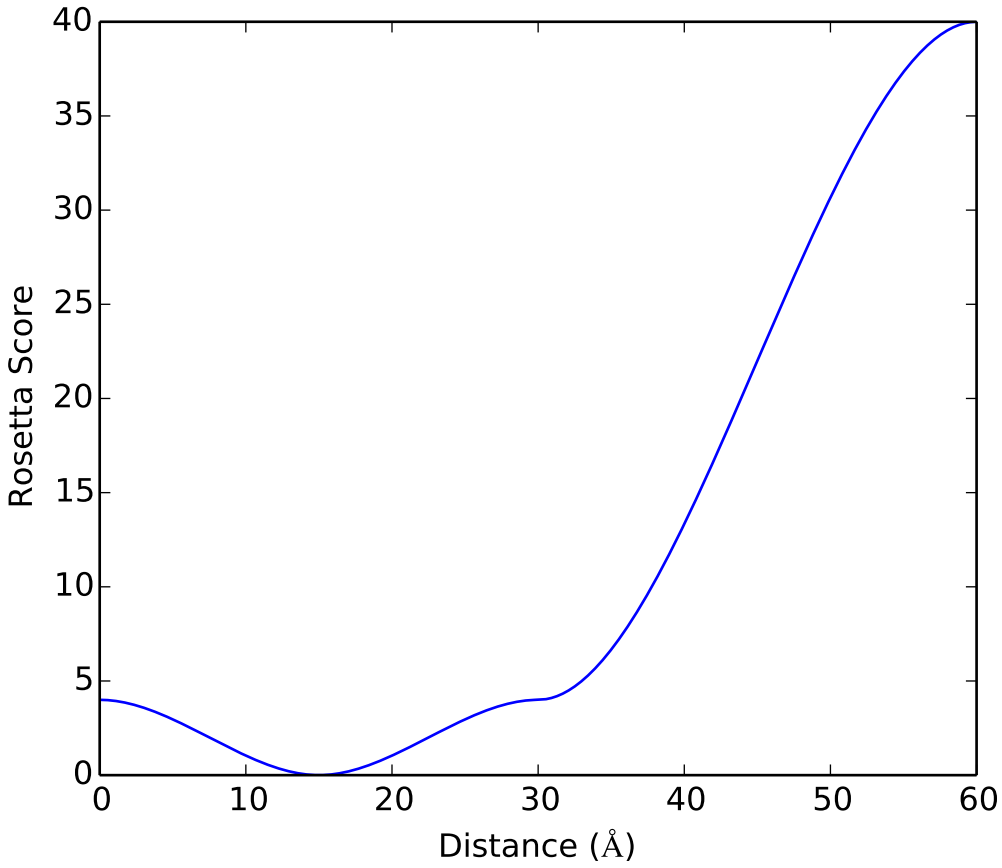
Supplementary Figure 4. Alternative nucleotide attachment sites for the radical source



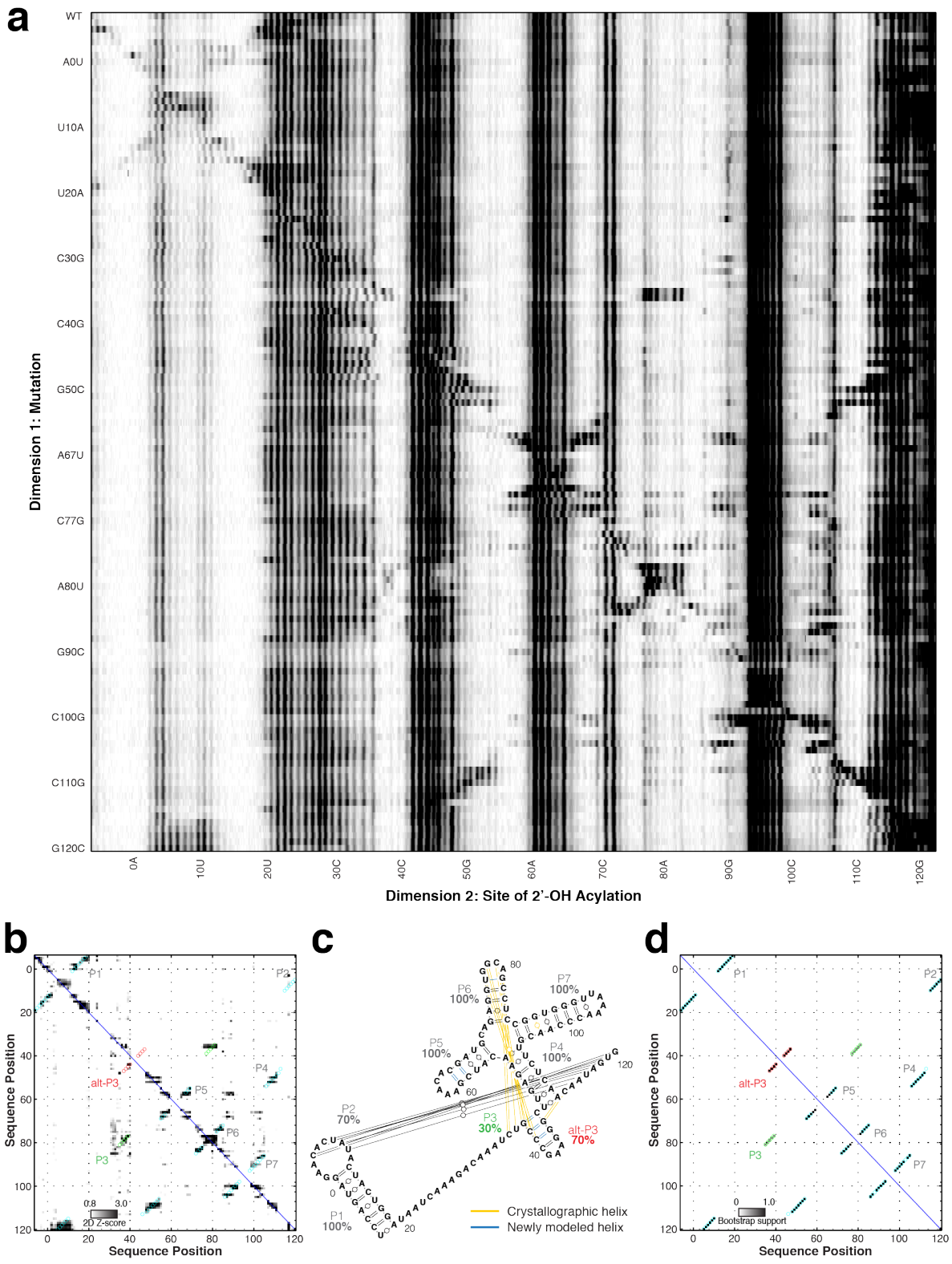
Supplementary Figure 5. Variation of fragmentation reaction time



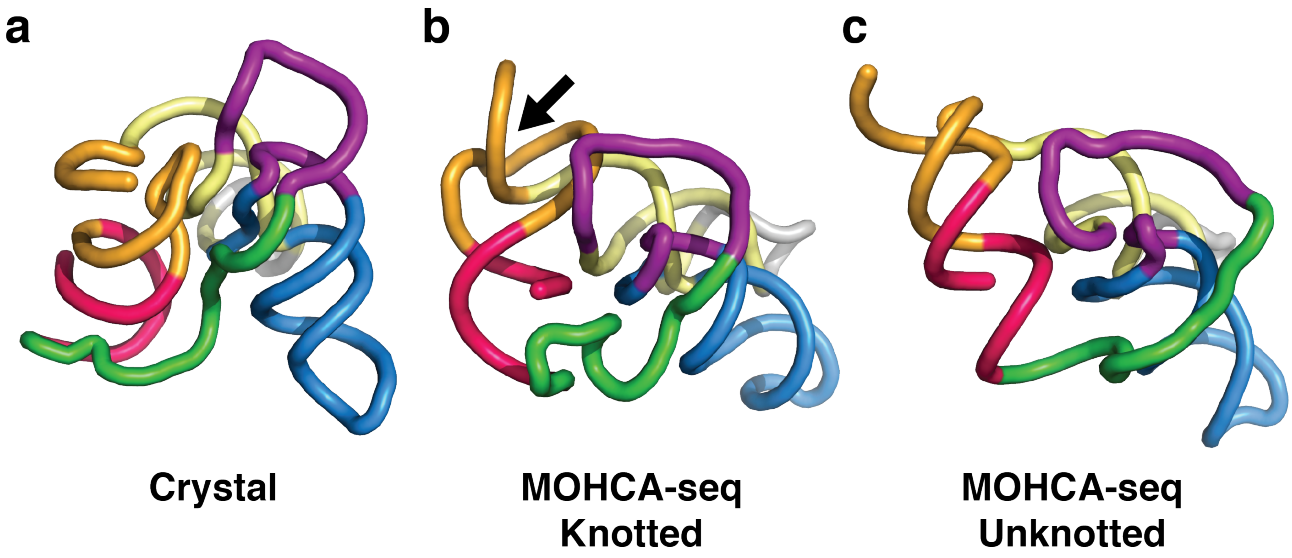
Supplementary Figure 6. Rosetta scoring potential used to incorporate MOHCA-seq constraints in modeling



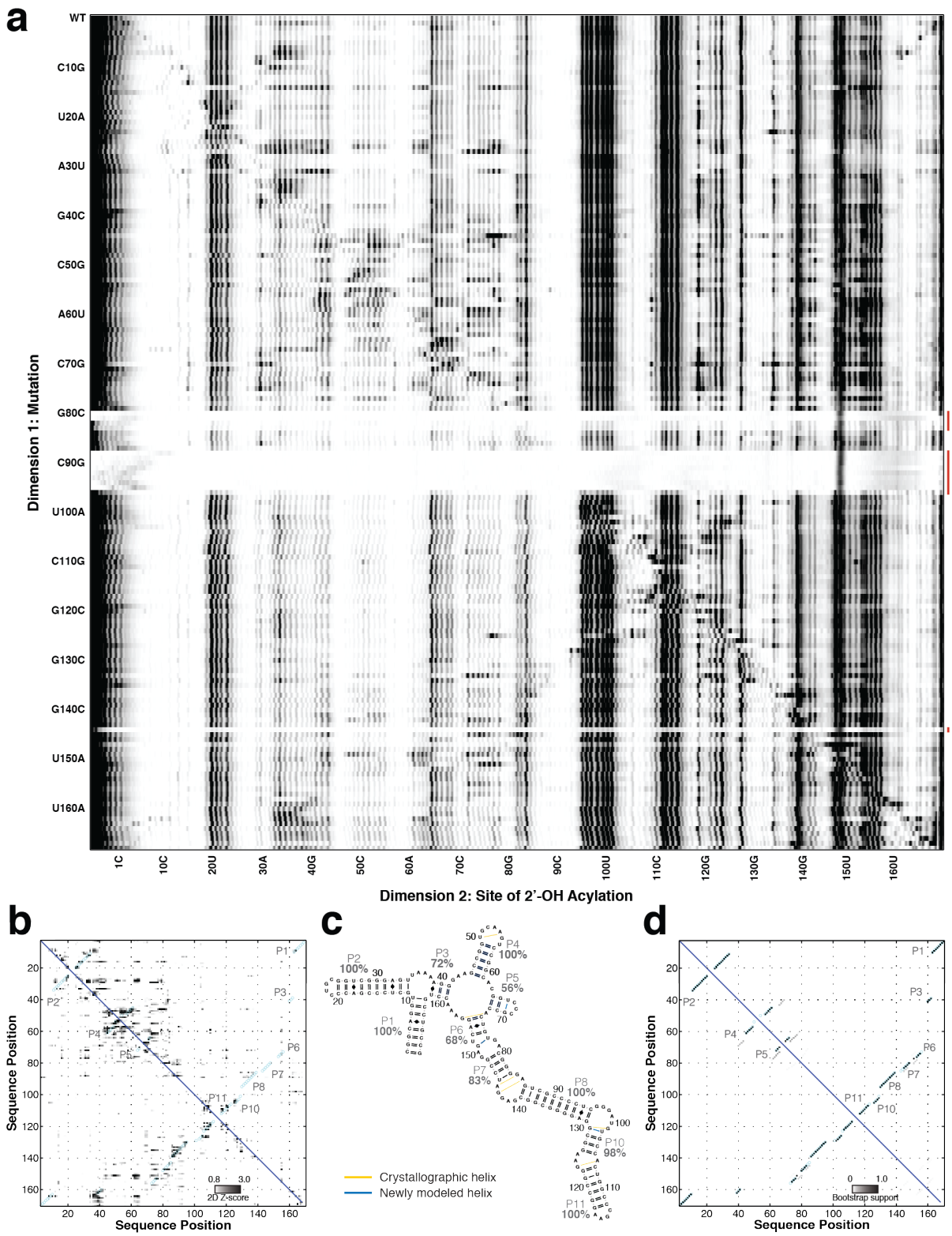
Supplementary Figure 7. Mutate-and-map analysis of class I ligase



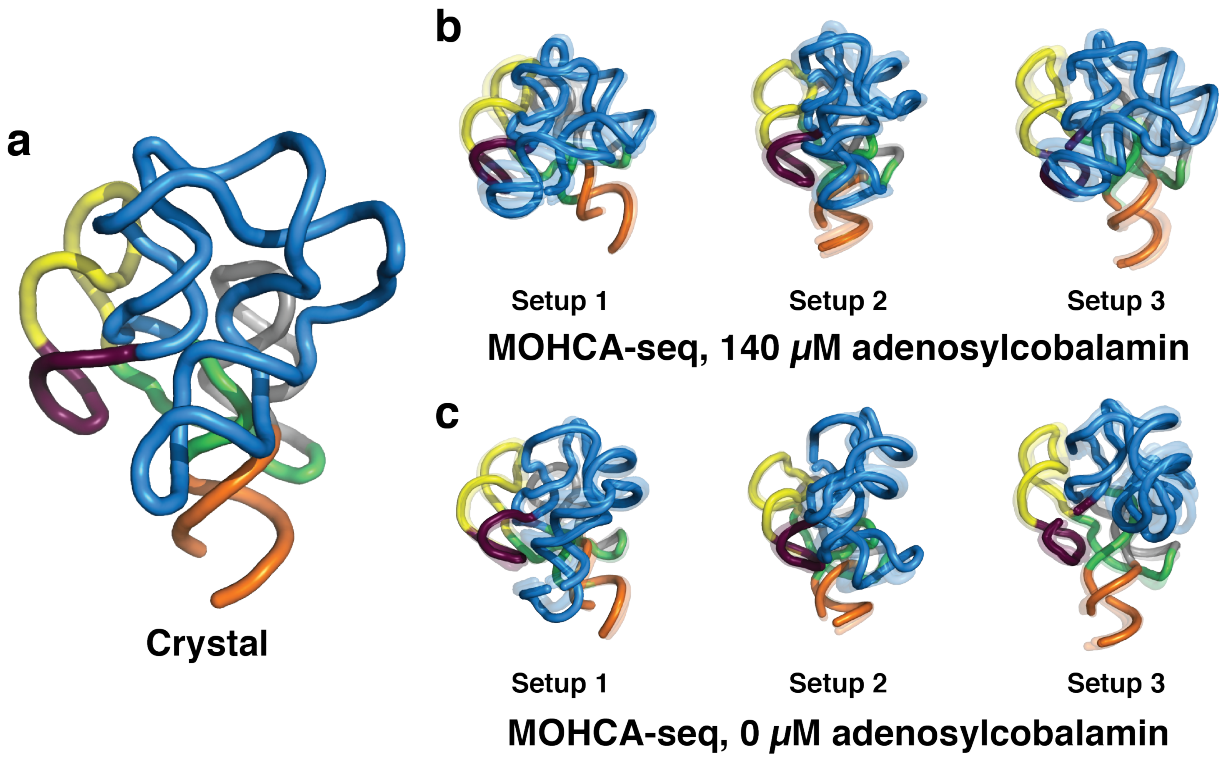
Supplementary Figure 8. Comparison of class I ligase crystal structure and knotted and unknotted cluster centers.

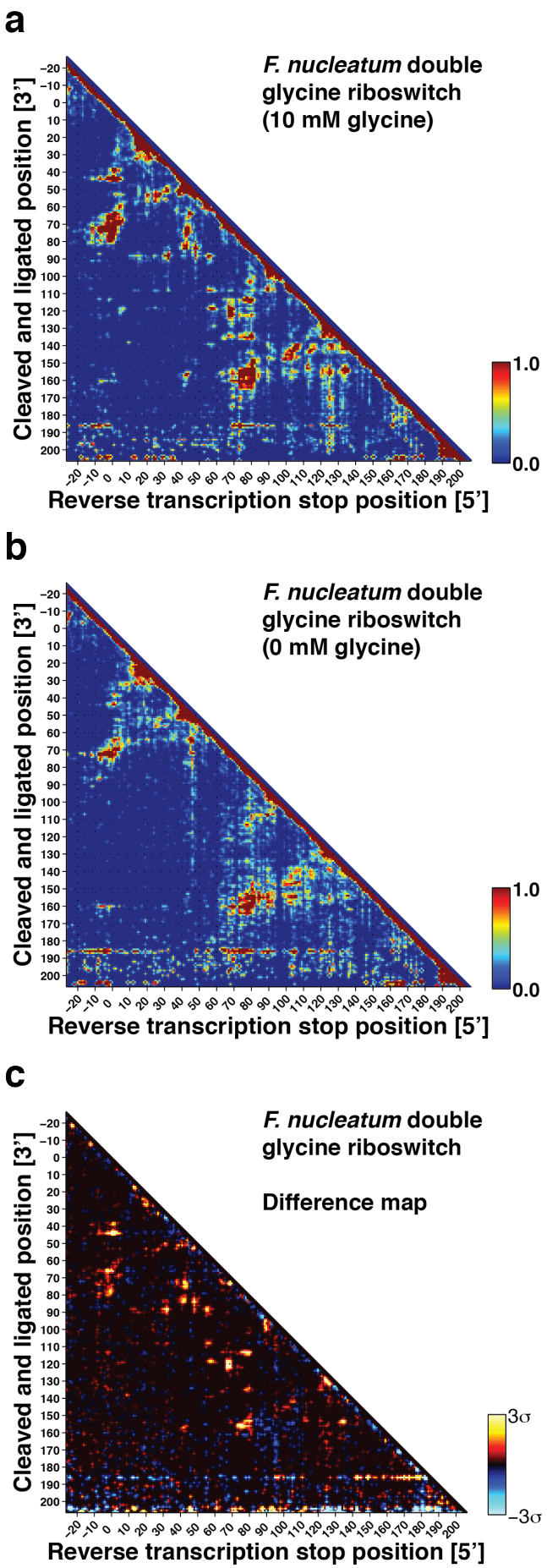


Supplementary Figure 9. Mutate-and-map analysis of the adenosylcobalamin riboswitch, the sixth target of the RNA-puzzles structure prediction trials

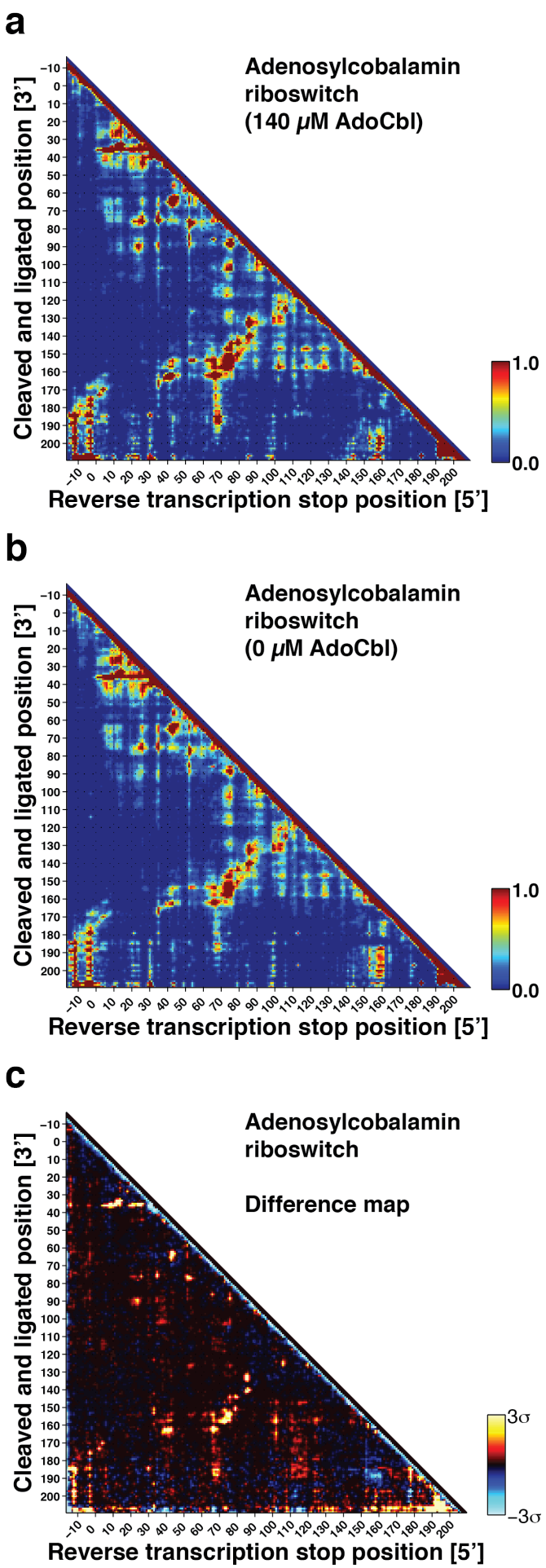


Supplementary Figure 10. Comparison between MOHCA-seq-guided modeling of the adenosylcobalamin riboswitch using different initial RNA fragment sets.





Supplementary Figure 11. Difference map comparison of the ligand-bound and ligand-free states of the glycine riboswitch.



Supplementary Figure 12. Difference map comparison of the ligand-bound and ligand-free states of the adenosylcobalamin riboswitch.

Supplementary Tables 1–3

Supplementary Table 1. Sequences of RNAs, single-stranded DNA ligation adapters, and sequencing primers

RNAs were transcribed *in vitro* from PCR-assembled DNA templates which included a promoter sequence (TTCTAATACGACTCACTATA) on the 5′ end for T7 RNA polymerase. Red = 5'-buffer region; black = region of interest; blue = 3'-buffer and tail region. 'RTU' sequences are reverse transcription primers that anneal to the universal miRNA cloning linker. Purple = 12-nt sequence-balanced barcode.

Name	Sequence
Tetrahymena ribozyme P4-P6	GGCCAAAACAACGGAAUUGCGGGAAAGGGGUCAACAGCCGUUCAGUACCAAGUCUCAGGGGAAACUUUGAGAUGGCCUUGCAAAGGGUAUGGUAAUAAGCUGACGGACAUGGUCCUAACCACGCAGCCAAGUCCUAAGUCAACAGAUCUUCUGUUGAUAUGGAUGCAGUUCAAAACCAAACCAAAGAAACAACAACAACAAC
Tetrahymena ribozyme P4-P6 with flanking hairpins	GGCCAAAGGCGUCGAGUAGACGCCAACAACGGAAUUGCGGGAAAGGGGUCAACAGCCGUUCAGUACCAAGUCUCAGGGGAAACUUUGAGAUGGCCUUGCAAAGGGUAUGGUAAUAAGCUGACGGACAUGGUCCUAACCACGCAGCCAAGUCCUAAGUCAACAGAUCUUCUGUUGAUAUGGAUGCAGUUCAAAACCAAACCGUCAGCGAGUAGCUGACAAAAAGAAACAACAACAACAAC
F. nucleatum double glycine riboswitch	GGCAAUUCGAGUAGAAUUGACAGAGAGGAUAUGAGGAGAGAUUUCAUUUUAAUGAAACACCGAAGAAGUAAAUCUUUCAGGUAAAAAGGACUCAUAUUGGACGAACCUCUGGAGAGCUUAUCUAAGAGAUAACACCGAAGGAGCAAAGCUAAUUUUAGCCUAAACUCUCAGGUAAAAGGACGGAGAAAACACAAGUUCAGGAGUACUGAACCAAAGAAACAACAACAACAAC
S. thermophilum adenosylcobalamin riboswitch	GGAACAGCCCGAGUAGGGCCGGCAGGUGCUCCCGACCCUGCGGUCGGGAGUUAAAAGGGAAGCCGGUGCAAGUCCGGCACGGUCCCGCCACUGUGACGGGGAGUCGCCCCUCGGGAUGUGCCACUGGCCCGAAGGCCGGGAAGGCGGAGGGGCGGCGAGGAUCCGGAGUCAGGAAACCUGCCUGCCGGCGCGAGUAGCGCAAACGAAAGAAACAACAACAACAAC
Class I Ligase	GGAACCGCGAGUAGCGGAAAUCCAGUAGGAACACUAUACUACUGGAUAAUCAAAGACAAAUCUGCCCGAAGGGCUUGAGAACAUCGAAACACGAUGCAGAGGUGGCAGCCUCCGGUGGGUUAAAACCCAACGUUCUCAACAAUAGUGAAAAGCGCGAGUAGCGCAACAAAGAAACAACAACAACAAC
Universal miRNA cloning linker (NEB)	rAppCTGTAGGCACCATCAAT–NH_2
Second ligation adapter	pAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGp
RTU048	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTAGCATGCTTAATTGATGGTGCCTACAG
RTU049	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGCTGCTACACATTGATGGTGCCTACAG
RTU050	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCAATACTGGGATTGATGGTGCCTACAG
RTU051	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTTGCGAGACTATTGATGGTGCCTACAG
RTU101	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTTTTCCCGCAGGATTGATGGTGCCTACAG
RTU102	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGAGGATTATACATTGATGGTGCCTACAG
RTU103	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCCGCATGATGCTATTGATGGTGCCTACAG
RTU104	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGACATGACGCTAATTGATGGTGCCTACAG
RTU105	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCCCCTTTTATACATTGATGGTGCCTACAG
RTU106	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTTTGCGACTGCTATTGATGGTGCCTACAG
RTU107	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGAGAACGGGCTAATTGATGGTGCCTACAG
RTU108	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGATGACACAGGATTGATGGTGCCTACAG

Supplementary Table 2. Pairwise MOHCA-seq constraints used for *de novo* modeling

***Tetrahymena* ribozyme P4-P6**	**5´-end**	**3´-end**	***F. nucleatum* double glycine riboswitch (10 mM glycine)**	**5´-end**	**3´-end**	***F. nucleatum* double glycine riboswitch (0 mM glycine)**	**5´-end**	**3´-end**
Strong	126	188	**Strong**	2	38	**Weak**	2	38
	123	194		1	44		11	26
	132	178		5	60		2	64
	130	166		2	64		5	57
	132	162		25	54		25	57
	132	159		45	64		45	64
	137	156		45	75		78	113
	136	162		32	88		86	108
	154	178		42	84		136	154
	171	211		48	84		100	148
	170	198		55	88		105	154
	170	190		55	108		74	157
	170	178		58	118			
	185	196		67	119			
	185	212		67	121			
	189	211		78	113			
	189	198		78	135			
	172	221		42	157			
	239	247		74	156			
	208	259		100	148			
	209	254		100	145			
	215	248		113	153			
	221	244		135	154			
	227	238		5	119			
	221	232	**Weak**	25	88			
	221	228		37	62			
Weak	189	224		79	103			
	202	214		15	88			
	114	168		32	108			

	189	224		9	138			
	202	214		25	118			
	151	225						
	108	254						
	154	254						
	189	255						
***S. thermophilum* adenosylcobalamin riboswitch (140 *μ*M adenosylcobalamin)**	**5´-end**	**3´-end**	***S. thermophilum* adenosylcobalamin riboswitch (0 *μ*M adenosylcobalamin)**	**5´-end**	**3´-end**	**Class I ligase**	**5´-end**	**3´-end**
Strong	11	26	**Strong**	12	25	**Strong**	6	41
	26	44		9	26		10	40
	5	42		26	42		31	84
	26	75		26	76		42	109
	35	75		42	56		42	116
	53	77		52	77		54	79
	35	89		76	88		63	88
	24	89		75	103		68	89
	52	89		85	133		67	94
	52	64		103	121		63	83
	75	88		35	167		70	103
	74	103		53	153		64	103
	73	117		66	154		64	109
	90	100		67	162		78	109
	85	132		74	147		80	103
	53	154		79	144	**Weak**	5	49
	67	153		86	133		5	110
	67	162		150	161		27	48
	73	147	**Weak**	6	41		35	70
	79	142		35	88		47	87
	102	147		75	117		47	103
	102	157		75	132		47	70
	118	147		90	107		76	113

	128	158		6	162		76	116
	146	158		43	154			
Weak	14	90		46	162			
	35	102		67	158			
	59	76		104	147			
	56	64		104	157			
	102	121		127	158			
	54	126		127	147			
	52	144		118	147			
	65	147		158	158			
	5	162						
	43	154						
	35	166						
	118	157						
	127	147						

Supplementary Table 3. Results of *de novo* modeling incorporating MOHCA-seq constraints

	# of models	Size of largest cluster	RMSD, cluster center (Å)	RMSD, cluster median (Å) (accuracy)	Clustering threshold (Å) (precision)	Total constraints	Percent strong constraints satisfied[a], models	Percent strong constraints satisfied, crystal[b]	Percent weak constraints satisfied, models	Percent weak constraints satisfied, crystal[b]
Tetrahymena ribozyme P4-P6	61,115	48	8.6	12.3	11.7	35	79.3	69.2	51.4	44.4
F. nucleatum double glycine riboswitch, bound	16,506	14	7.9	10.3	7.1	31	93.5	100	70.4	42.9
F. nucleatum double glycine riboswitch, unbound[b]	15,771	12	25.4	17.7	18.9	12	N/A[c]	N/A[c]	64.6	83.3
Class I ligase (unknotted)	17,881	7	14.5	15.4	8.2	24	81.9	53.3	44.4	77.8
Class I ligase (knotted)	17,881	14	16.1	15.5	8.2	24	85.2	53.3	61.9	77.8
S. thermophilum adenosylcobalamin riboswitch, bound[d]	14,219	12	12.1	12.4	9.8	38	84.7	76.0	85.3	69.2
S. thermophilum adenosylcobalamin riboswitch, unbound[b,d]	11,980	10	17.3	15.6	13.9	33	82.8	94.4	82.7	73.3

[a] Constraints were considered satisfied if the O2′ of the 5′-residue was less than 30 Å from the C4′ of the 3′-residue (Supplementary Table 2).
[b] For ligand-free states of the glycine riboswitch and adenosylcobalamin riboswitch, RMSDs and percent constraints satisfied were calculated for gold standard crystal structures solved in the presence of ligand.
[c] No strong constraints were selected for the unbound state of the glycine riboswitch.
[d] (See description of AdoCbl riboswitch modeling in Methods.) Three separate modeling runs were performed for each ligand-binding state of the AdoCbl riboswitch. The number of models, size of largest cluster, cluster center RMSD, median Rosetta energy score, and percent constraints satisfied statistics for the AdoCbl riboswitch are representative data from the modeling runs with the most models generated. The cluster median (accuracy) shown is the median of the three cluster median RMSDs of the largest clusters generated in each of the three modeling runs. The clustering threshold (precision) shown is calculated as the mean pairwise RMSD between the cluster centers of the largest clusters generated in each of the three modeling runs.