Supplementary Material for

Inferring RNA structure from mobility-based deep mutational landscapes

Jinle Tang[1†], Yazhou Shi[1,2†], Zhe Zhang[1], Dailin Luo[1,3], Jian Zhan[1,4*], and Yaoqi Zhou[1*]

[1]Institute of Physical and Systems Biology, Shenzhen Bay Laboratory, Shenzhen, Guangdong Province, 518107, China

[2]Research Center of Nonlinear Science, School of Mathematics & Statistics, Wuhan Textile University, Wuhan, 430200, China.

[3]School of Life and Health Sciences, The Chinese University of Hong Kong, Shenzhen, 518172, China

[4]Ribopeutic Inc., Guangzhou International Bio Island, Guangdong, 510320, China

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

* To whom correspondence should be addressed. Yaoqi Zhou. Tel: +86 755 26849275; Email: zhouyq@szbl.ac.cn; Jian Zhan. Tel: +86 755 26849280; Email: zhanjian@szbl.ac.cn.

Supplementary Table S1. The structural similarity according to TM-score between eight RNAs employed in this work.

RNAs	SAM-VI riboswitch	Adenine riboswitch	xrRNA	CPEB3 ribozyme	Cloverleaf RNA	SARS-CoV-2 SL5	RRE SLII	raiA
SAM-VI riboswitch	1.0	0.102	0.036	0.106	0.05	0.105	0.07	-
Adenine riboswitch	0.086	1.0	0.044	- [a]	-	-	-	-
xrRNA	0.028	0.042	1.0	0.042	0.014	0.028	0.068	0.083
CPEB3 ribozyme	0.094	-	0.044	1.0	0.077	-	0.084	-
Cloverleaf RNA	0.02	-	0.007	0.044	1.0	0.027	0.021	-
SARS-CoV-2 SL5	0.068	-	0.016	-	0.032	1.0	0.032	-
RRE SLII	0.029	-	0.007	0.048	0.023	0.03	1.0	0.007
raiA	-	-	0.035	-	-	-	0.005	1.0

[a] "-" represents that there is no alignment between two RNAs.

Supplementary Table S2. The DNA template sequences of native RNAs as well as the designed double mutations (denoted as 2M), and the vector pUC57-T7Q-RNA.

Name [a]		Sequence (5'-3') [b]
SAM-VI riboswitch	WT	GGCATTGTGCCTCGCATTGCACTCCGCGGGGCGATAAGTCCTGAAAAGGGATGTC
	2M	GGCATTGTGCCTCGCATTGCACTCCGCTGGGCGATAAGTCCGGAAAAGGGATGTC
Adenine riboswitch	WT	GGGAAGATATAATCCTAATGATATGGTTTGGGAGTTTCTACCAAGAGCCTTAAACTCTTGATTATCTTCCC
	2M	GGGAAGATATAATACTAATGATATGGTTTGGGAGTTTCTACCTAGAGCCTTAAACTCTTGATTATCTTCCC
xrRNA	WT	GGGTCAGGCCGGCGAAAGTCGCCACAGTTTGGGGAAAGCTGTGCAGCCTGTAACCCCCCCACGAAAGTGGG
	2M	GGGTCACGCCGGCGAAAGTCGCCAGAGTTTGGGGAAAGCTGTGCAGCCTGTAACCCCCCCACGAAAGTGGG
CPEB3 ribozyme	WT ($C_{57}T$)	TAACAGGGGGCCACAGCAGAAGCGTTCACGTCGCAGCCCCTGTCAGATTCTGGTGAATCTGTGAATTCTGCTGTATATCTC
	2M	TAACAGGGGGCCACATCAGAAGCGTTCACGTCGCAGCCCCTGTCAAATTCTGGTGAATCTGTGAATTCTGCTGTATATCTC
Cloverleaf RNA (R1116)	WT	CGCCCGGATAGCTCAGTCGGTAGAGCAGCGGCTAAAACAGCTCTGGGGTTGTACCCACCCCAGAGGCCCACGTGGCGGCTAGTACTCCGGTATTGCGGTACCCTTGTACGCCTGTTTTAGCCGCGGGTCCAGGGTTCAAGTCCCTGTTCGGGCGCCA
	2M	CGCGCGGATAGCTCAGTCGGTAGAGCAGCGGCTAAAACAGCTCTGGGGTTGTACCCACGCCAGAGGCCCACGTGGCGGCTAGTACTCCGGTATTGCGGTACCCTTGTACGCCTGTTTTAGCCGCGGGTCCAGGGTTCAAGTCCCTGTTCGGGCGCCA
SARS-CoV-2 SL5 (R1149)	WT	GGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCC
	2M	GGACACGAGTAACTCGTCTATCTTCTGCAGGCAGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATGTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCC
RRE SLII（R1203）	WT	GCCCGGATAGCTCAGTCGGTAGAGCAGCGGGCACTATGGGCGCAGTGTCAATGGACGCTGACGGTACAGGCCAGACAATTATTGTCTGGTATAGTGCCCGCGGGTCCAGGGTTCAAGTCCCTGTTCGGGCGCCA
raiA（R1242）	WT	GGTTAAGTTAGGTTTGTGGTTGAAAGTCGATGCCAGTCGCAGGCAAAACGATCCACGTAAGTTAAACAAAGTTTTAATGAGCATGGTGCGGCTTAGAAGTAAGTCCTGCCGCTTTAGGCGAGAGTATTAGTAGTGAGAGGGTAATTCCGGGTAGCGAAACTTCCAGCAGGCGAGTGTGGGGTCAAAGACCAGGTCAACTAACTTA
PZ50	WT	GGAACCTCCGCCGAAAGGCGGTGAAGGAGAGGCGCAAGGTTAACCGCCTCAGGTTCC
PZ51	WT	GGCGGCCCCTGAATGCGGCTAATCCTAACTGCGGAGCACACACCCTCGAAACACGAGGGCAGTGTGTCGTAACGGGCAACTCTGCAGCGGAACCGACTAC

		TTTGGGTGTCCGCC
PZ62	WT	GGAGCGTAATCGCGTCTTGACGTGAGTGAGCCTAGACTGTAGGCTATCCCT TCGGGGATAGCATAGCTCACAATACGCTGATCTACAGTCA
pUC57-T7Q-RNA		TAATACGACTCACTATAGCGGTTTTCCCAGTCACGACXXXXXXCACTGGC CGTCGTTTTACAGAAGAGCACATGTGAGCAAAAGGCCAGCAAAAGGCCA GGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCC CCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACC CGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTG CGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTC CCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGT TCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGT TCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCC GGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTA GCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCT AACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAA GCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAA CCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGC AGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGA CGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTAT CAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAA TCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTT AATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGT TGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCAT CTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCA GATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTG GTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAG CTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTG CTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCT CCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAA AAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGC CGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGT CATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTC ATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAA TACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATT GGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAG ATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTT TACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCG CAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTT CCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGG ATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCA CATTTCCCCGAAAAGTGCCACCTGACGTCGATATC

[a] "WT" represents the DNA template for the native RNA sequence; "2M" represents the DNA template for a disruptive double mutant shown in Figure 2 in the main text. [b] The mutant bases are marked with red color. For CPEB3 ribozyme, in order to facilitate mobility screening, the 57-th base (i.e., C) was also mutated into T (marked with green color) to reduce their activity. For pUC57-T7Q-RNA, "XXXXXX" represents the DNA template sequences of native RNAs. The T7 RNA polymerase promoter sequence, adaptor sequence, and BspQ I recognition sequence are highlighted in brown, cyan, and purple, respectively.

Supplementary Table S3. The number of variants and the coverage of single, double, triple, and other mutations of eleven RNAs examined.

RNAs	Length (nt)	Read counts (wild type) [a]	Mutations [b]	Number of variants	Mutation coverage [c]
SAM-VI riboswitch	55	5863353 (5745118)	1	165	100%
			2	12722	95.2%
			3	24966	3.5%
			4-9	69	-
Adenine riboswitch	71	4369695 (3763801)	1	213	100%
			2	19770	88.4%
			3	4546	0.29%
			4-8	86	-
xrRNA	71	8203434 (4400848)	1	213	100%
			2	21121	94.4%
			3	89575	5.8%
			4-10	38020	-
CPEB3 ribozyme	81	6155432 (2465101)	1	243	100%
			2	26741	91.7%
			3	66277	2.9%
			4-10	28032	-
Cloverleaf RNA	157	6077690 (5157983)	1	471	100%
			2	90702	82.3%
			3	47272	0.28%
			4-10	19995	-
Cloverleaf RNA (Oligo-pool)	157	2659405 (1597572)	1	471	100%
			2	102801	93.3%
			3	13611	0.08%
			4-10	1591	-
SARS-Cov-2 SL5	124	6503023 (6051484)	1	372	100%
			2	63109	92.0%
			3	199707	2.4%
			4-10	133955	-
SARS-Cov-2 SL5 (Oligo-pool)	124	3447467 (2266332)	1	372	100%
			2	68285	99.5%
			3	6477	0.08%
			4-10	580	-

RRE SLII	134	19422661 (15671699)	1	402	100%
			2	70416	87.8%
			3	130381	1.2%
			4-10	56467	-
raiA	205	11240707 (2892238)	1	615	100%
			2	152373	81.0%
			3	146258	0.4%
			4-10	169898	-
PZ50	57	21351459 (16462523)	1	171	100%
			2	14047	97.8%
			3	93907	11.9%
			4-10	34838	-
PZ51	114	21352075 (17173866)	1	342	100%
			2	52871	91.2%
			3	145128	2.2%
			4-10	19454	-
PZ62	91	6130578 (5098261)	1	273	100%
			2	33610	91.2%
			3	127892	3.9%
			4-10	11292	-

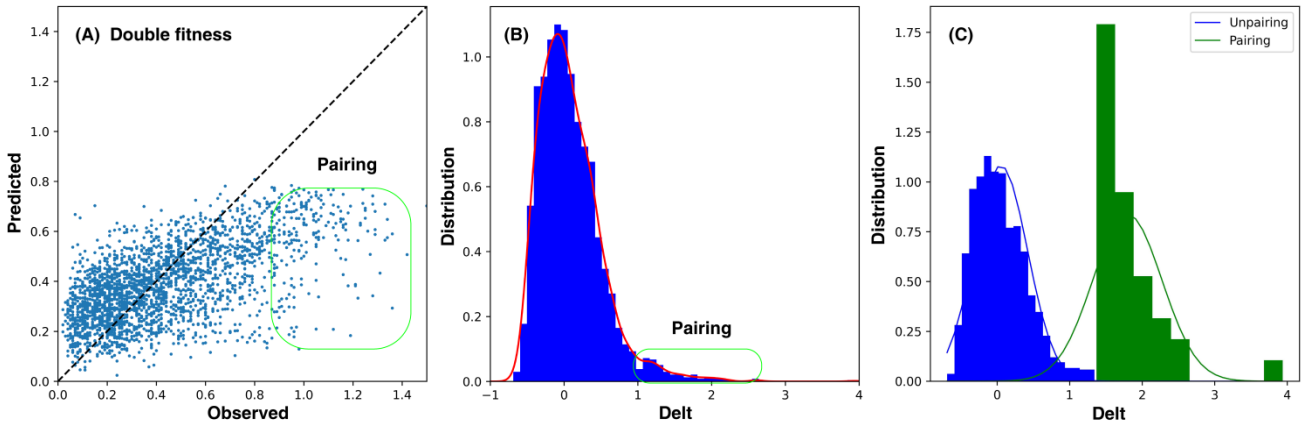
[a] The read count (the corresponding wild-type reads) for each library. For Cloverleaf RNA and SARS-CoV-2 SL5, the read counts from the oligonucleotide pool libraries are shown in a separate row. [b] The number of mutated bases in each sequence. Because the sequences with multiple mutations are relatively fewer (as we were more interested in double mutations), we have consolidated multiple mutations (>3) together. [c] "-" indicates that the mutation coverage is very low (~0%).

Supplementary Table S4. Comparison of Matthews Correlation Coefficient (MCC) inferred by CODA, the RNA-specific CODA (i.e., CODA2), and signal amplification by Monte Carlo simulations (CODA2 + MC, the best in top 5) and further by BRiQ (CODA2 + MC + BRiQ, the best in top 5), as well as R-scape, the mean-field direct coupling analysis (mfDCA-RNA), pseudolikelihood maximization coupling analysis (plmDCA-RNA), epistasis from Schmiedel & Lehner (epistasis-SL) and from Rollins et al (epistasis-Rollins).

Methods \ RNA	SAM-VI riboswitch	Adenine riboswitch	xrRNA	CPEB3 ribozyme	Cloverleaf RNA	SARS-CoV-2 SL5	RRE SLII	raiA
mfDCA	0.15	0.13	0.20	0.12	0.14	0.11	0.12	0.19
plmDCA	0.16	0.15	0.21	0.13	0.15	0.10	0.10	0.22
epistasis-Rollins	0.11	0.23	0.15	0.11	0.13	0.12	0.10	0.15
epistasis-SL	0.15	0.21	0.18	0.13	0.12	0.12	0.10	0.12
R-scape	0.12	0.32	0.41	0.21	0.16	0.13	0.16	0.27
CODA	0.19	0.33	0.48	0.25	0.18	0.13	0.16	0.34
CODA2	0.34	0.53	0.64	0.40	0.42	0.25	0.25	0.55
CODA2+MC	0.85	0.89	0.91	0.86	0.91	0.96	0.88	0.90
CODA2+MC+BRiQ	0.90	0.92	0.94	0.92	0.92	0..93	0.92	0.91

Supplementary Table S5. The primers used in this study.

Name	Sequence (5' - 3')	Notes
M13-FP	GGTTTTCCCAGTCACGAC	EP-PCR primer
M13-RP	GTAAAACGACGGCCAGTG	EP-PCR primer
M13-Plus-FP	TAGCGGTTTTCCCAGTCACGAC	Amplification of EP-PCR products for library construction
M13-Plus-RP	CTTCTGTAAAACGACGGCCAGTG	Amplification of EP-PCR products for library construction
AS-FP	CACTGGCCGTCGTTTTACAGAAGAGCacat	Amplification of vector for library construction
AS-RP	GTCGTGACTGGGAAAACcgctatagtgagtcgtattag	Amplification of vector for library construction
P5-089-FP	AATGATACGGCGACCACCGAGATCTACAC ACACTAAGACACTCTTTCCCTACACGACGC TCTTCCGATCTgGTTTTCCCAGTCACGAC	For NGS Samples preparation
P5-090-FP	AATGATACGGCGACCACCGAGATCTACAC GTGTCGGAACACTCTTTCCCTACACGACGC TCTTCCGATCTgGTTTTCCCAGTCACGAC	For NGS Samples preparation
P5-091-FP	AATGATACGGCGACCACCGAGATCTACACT TCCTGTTACACTCTTTCCCTACACGACGCTC TTCCGATCTgGTTTTCCCAGTCACGAC	For NGS Samples preparation
P5-092-FP	AATGATACGGCGACCACCGAGATCTACACC CTTCACCACACTCTTTCCCTACACGACGCTC TTCCGATCTgGTTTTCCCAGTCACGAC	For NGS Samples preparation
P7-089-RP	CAAGCAGAAGACGGCATACGAGATGTGCG ATAGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCTGTAAAACGACGGCCAGTG	For NGS Samples preparation
P7-090-RP	CAAGCAGAAGACGGCATACGAGATACATA GCGGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCTGTAAAACGACGGCCAGTG	For NGS Samples preparation
P7-091-RP	CAAGCAGAAGACGGCATACGAGATGAACA TACGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCTGTAAAACGACGGCCAGTG	For NGS Samples preparation
P7-092-RP	CAAGCAGAAGACGGCATACGAGATAGGTG CGTGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCTGTAAAACGACGGCCAGTG	For NGS Samples preparation



Supplementary Figure S1. A schematic principle of classification in CODA2. (A) The fitness of double mutants observed from sequencing versus those predicted by the independent-mutation model for xrRNA. Outliers for likely base pairs (co-variated) are indicated by the green box. (B) The distribution (blue bar) as well as the kernel density estimation curve (red line) of the difference (i.e., Delt) between observed and predicted values of double fitness, and the distribution of possible base pairs is marked by the green box. (C) The data of Delt shown in (B) can be divided into two contributions representing whether they are likely to form pairs, according to whether the data is more than $n * \sigma$ (σ is the standard deviation of data, and n is 3.5 here), and the distributions of these two sets of data along with their corresponding normal distribution curves are shown here.