Evaluation of novel computational methods to identify RNA-binding protein footprints from structural data

SUPPLEMENTAL MATERIAL

Orel Mizrahi[1,2,3,*], Meredith Corley[1,*], Ori Feldman[4,*], Thorben Fröhlking[5,6,7,8,*], Lei Sun[9,10,11,12,*], Alison Ziesel[13,*], Maciej Antczak[14,15], Mattia Bernetti[16,17,8], Shaimae I. Elhajjajy[18], Wenze Huang[12,19], Grady G. Nguyen[1,2,3], Samuel S. Park[1,2,3], Raul I. Perez Martell[20], Luke Trinity[20], Kui Xu[21], Tomasz Zok[14], Giovanni Bussi[8], Hosna Jabbari[13,22], Yaron Orenstein[23,24], Sharon Aviran[25,26], Michelle M. Meyer[27], Gene W. Yeo[1,2,3,28,29]

Corresponding authors: saviran@ucdavis.edu, m.meyer@bc.edu, geneyeo@ucsd.edu

* These authors contributed equally to this work.

Author Affiliations

1 Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA 92037, USA

2 Institute for Genomic Medicine, University of California San Diego, La Jolla, CA 92037, USA

3 Sanford Stem Cell Institute and Stem Cell Program, University of California San Diego, La Jolla, CA 92037, USA

4 School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer Sheva 8410501, Israel

5 School of Pharmaceutical Sciences, University of Geneva, 1206 Geneva, Switzerland

6 Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, 1206 Geneva, Switzerland

7 Swiss Institute of Bioinformatics, University of Geneva, 1206 Geneva, Switzerland

8 Scuola Internazionale Superiore di Studi Avanzati (SISSA), 34136 Trieste, Italy

9 Pingyuan Laboratory, Xinxiang 453007, Henan, China

10 Shandong Provincial Key Laboratory of Animal Cell and Developmental Biology, School of Life Sciences, Shandong University, Qingdao 266237, China

11 State Key Laboratory of Microbial Technology, Shandong University, Qingdao 266237, China

12 MOE Key Laboratory of Bioinformatics, Beijing Advanced Innovation Center for Structural Biology & Frontier Research Center for Biological Structure, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China

13 Department of Biomedical Engineering, University of Alberta, Edmonton, AB, Canada

14 Institute of Computing Science, Poznan University of Technology, 60-965 Poznan, Poland

15 Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

16 Computational and Chemical Biology, Fondazione Istituto Italiano di Tecnologia (IIT), I-16163 Genova, Italy

17 Department of Biomolecular Sciences, University of Urbino Carlo Bo, 61029 Urbino, Italy

18 Department of Genomics and Computational Biology, University of Massachusetts, Chan Medical School, Worcester, MA 01605, USA

19 Tsinghua-Peking Center for Life Sciences, Beijing 100084, China

20 Department of Computer Science, University of Victoria, Victoria, BC, Canada

21 State Key Laboratory of Membrane Biology, Beijing Frontier Research Center for Biological Structures, Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China

22 Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada

23 Department of Computer Science, Bar-Ilan University, Ramat Gan 5290002, Israel

24 The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 5290002, Israel

25 Department of Biomedical Engineering, University of California Davis, Davis, CA 95616, USA

26 Genome Center, University of California Davis, Davis, CA 95616, USA

27 Department of Biology, Boston College, Chestnut Hill, MA 02467, USA

28 Sanford Laboratories for Innovative Medicine, La Jolla, CA 92037, USA

29 Center for RNA Technologies and Therapeutics, La Jolla, CA 92037, USA

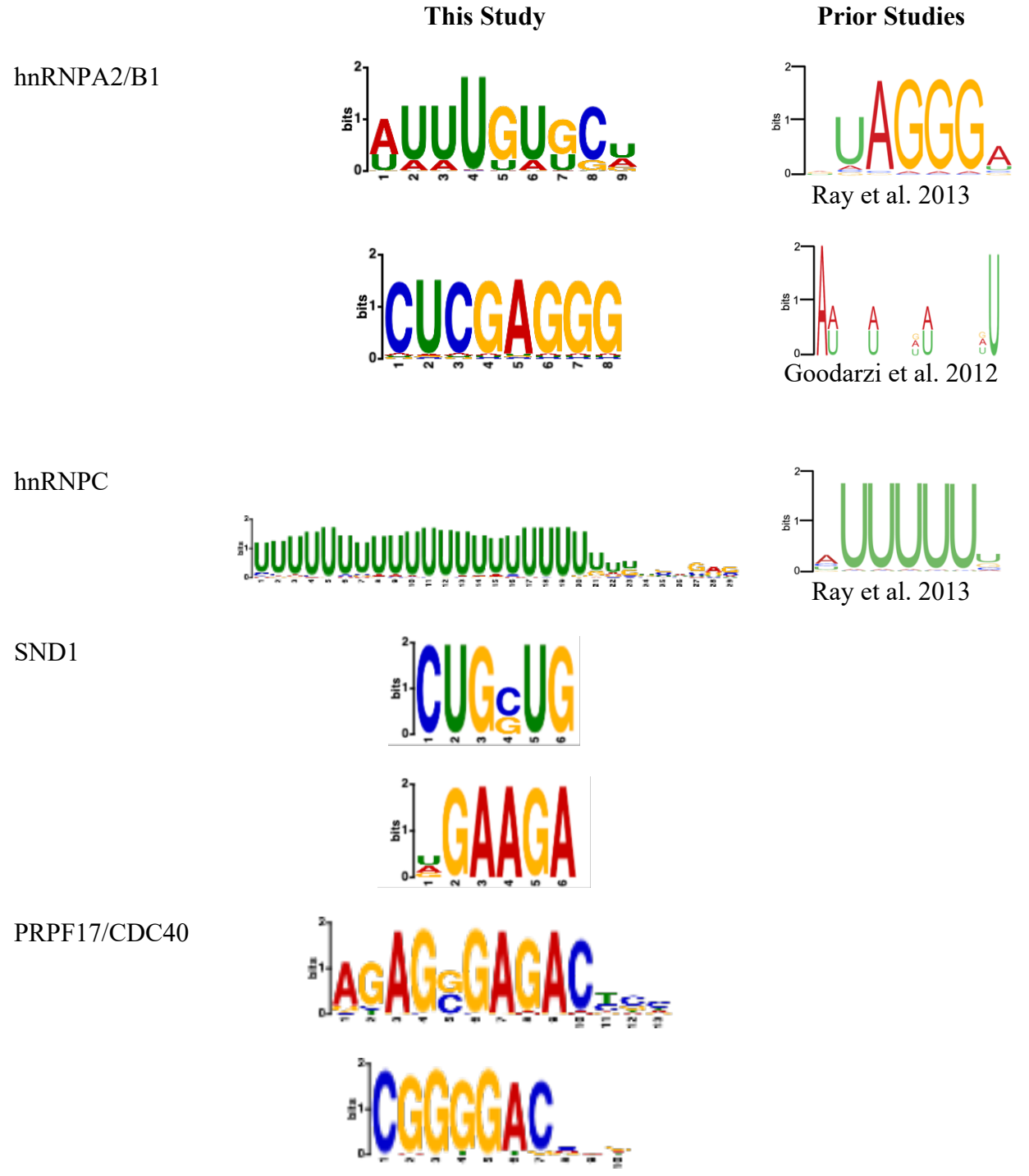
Supplemental Tables

Utility	Input	Prediction	Regions	Notes	Reference
KYG	structure	Yes	98 - 101 108 - 115 178 - 183 196 - 201 224 - 229 248 - 255 264 - 269 285 - 305 353 - 358 370 - 381 432 - 436 453 - 457 488 - 511 533 - 548		Kim et al. 2006
TriPepSVM	sequence	Yes	n/a	overall prediction	Bressin et al. 2019
SPOT-Struct-RNA	structure	No	n/a		Zhao et al. 2011
SPOT-Seq-RNA	sequence	Yes	288, 333, 357, 369		Yang et al. 2014
FastRNABindR	sequence	No	n/a		El-Manzalawy et al. 2016
HybridNAP	sequence	Yes	88 - 113 125 - 141 222 - 235		Zhang et al. 2019
DRNAPred	sequence	No	n/a		Yan and Kurgan 2017

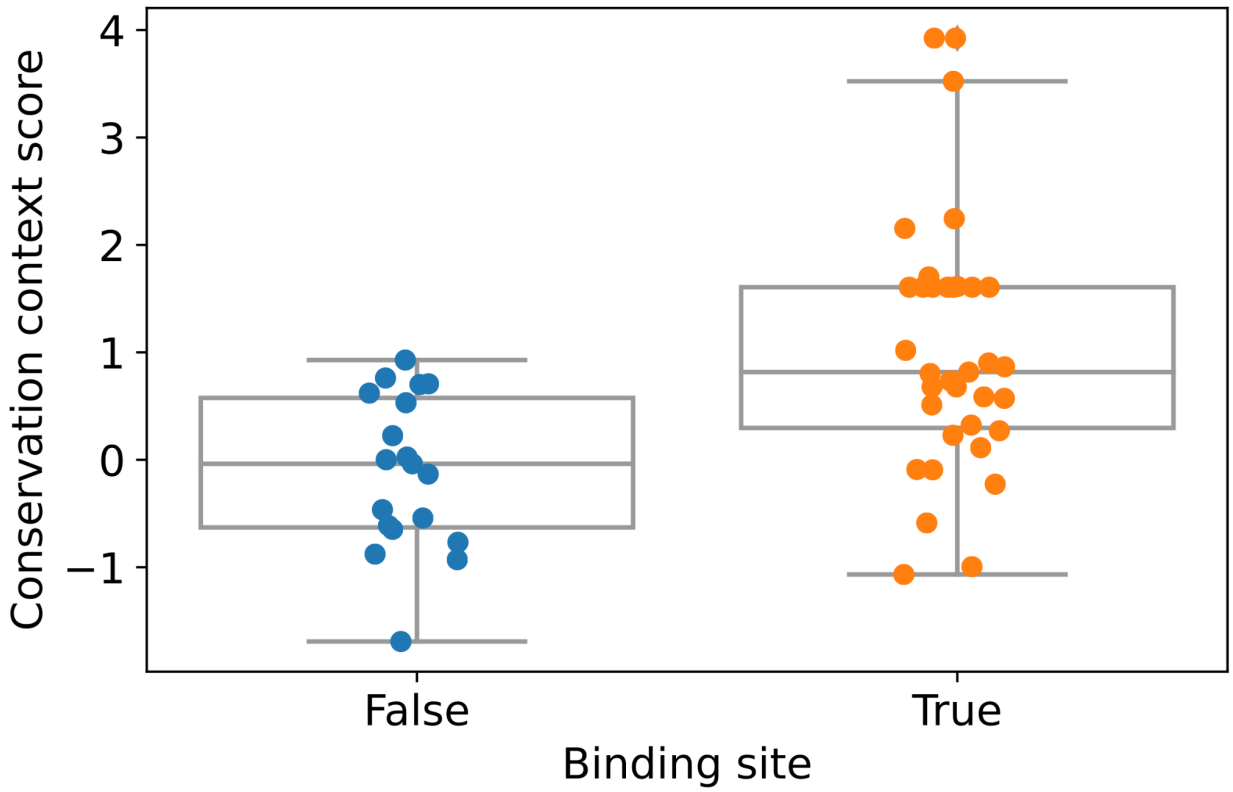
PPrint	sequence	Yes	16 - 21 169 - 184 225 - 231 255 - 265 281 - 289 305 - 311 376 - 383 401 - 412 437 - 450 480 - 493		Kumar et al. 2007
aaRNA	sequence	Yes	40 - 43 126 - 130 153 - 157 216 - 220 298 - 304		Li et al. 2014
Final Prediction	-	Yes	225-229		

Supplemental Table S1. Protein Binding Target Tool Glossary. First column lists utility name, second column states required input (protein sequence or structure), third column denotes if the utility positively predicted that PRPF17 can bind RNA, fourth column indicates predicted binding regions, and fifth column provides the relevant publication. For TriPepSVM, predictions are made at the protein rather than amino acid level, and we indicate this as 'overall prediction' in the Notes column. Overall, six utilities predicted that PRPF17 can bind RNA, but only five of them characterized specific residues as potentially capable of binding. If a region was identified by a majority of these utilities (i.e., three out of five), it was included in our final prediction (see bottom row). Final predictions were validated through protein-RNA specific docking with five RNAs selected for their relative abundance of identified RNA binding motifs in their 3' UTR sequence.

Supplemental Figures



Supplemental Figure S1. Sequence binding motifs found by each team in their top predicted sites and those found by prior studies. For each RBP, motif(s) mined by each team are shown on the left. If available, motif(s) found in prior studies are shown on the right (motifs were created to depict study results). References to these studies are listed below each motif.



Supplemental Figure S2. Conservation context scores in known iron response protein (IRP) binding sites ("True") are higher than in incorrect IRP binding sites ("False"), p-value = 0.0003. Context scores are calculated as the average PhyloP 100way score in the binding site subtracted by the average of the PhyloP 100way scores in flanking sites of the same size. Higher PhyloP 100way scores indicate higher conservation, and higher context scores indicate higher conservation in the binding site relative to its surrounding context. PhyloP 100way scores are available on the UCSC genome browser.

REFERENCES

Kim OTP, Yura K, Go N. 2006. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res* **34:** 6450-6460. doi:10.1093/nar/gkl819

Bressin A, Schulte-Sasse R, Figini D, Urdaneta EC, Beckmann BM, Marsico A. 2019. TriPepSVM: *de novo* prediction of RNA-binding proteins based on short amino acid motifs. *Nucleic Acids Res* **47:** 4406-4417. doi:10.1093/nar/gkz203

Zhao H, Yang Y, Zhou Y. 2011. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Res* **39:** 3017-3025. doi:10.1093/nar/gkq1266

Yang Y, Zhao H, Wang J, Zhou Y. 2014. SPOT-Seq-RNA: Predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction. In *Protein Structure Prediction* (ed. Kihara D), pp. 119-130. Humana Press, New York, NY. doi:10.1007/978-1-4939-0366-5_9

El-Manzalawy Y, Abbas M, Malluhi Q, Honavar V. 2016. FastRNABindR: Fast and accurate prediction of protein-RNA interface residues. *PLoS ONE* **11:** e0158445. doi:10.1371/journal.pone.0158445

Zhang J, Ma Z, Kurgan L. 2019. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief Bioinform* **20:** 1250-1268. doi:10.1093/bib/bbx168

Yan J, Kurgan L. 2017. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res* **45:** e84. doi:10.1093/nar/gkx059

Kumar M, Gromiha MM, Raghava GPS. 2008. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins*: *Structure, Function, and Bioinformatics* **71:** 189-194. doi:10.1002/prot.21677

Li S, Yamashita K, Amada KM, Standley DM. 2014. Quantifying sequenced and structural features of protein-RNA interactions. *Nucleic Acids Res* **42:** 10086-10098. doi:10.1093/nar/gku681

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499:** 172-177. doi:10.1038/nature12311

Goodarzi H, Najafabadi HS, Oikonomou P, Greco TM, Fish L, Salavati R, Cristea IM, Tavazoie S. 2012. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature* **485:** 264-268. doi:10.1038/nature11013