SUPPLEMENTARY METHODS FOR

*cat*RAPID *omics v2.0*: going deeper and wider in the prediction of protein–RNA interactions

Alexandros Armaos [1], Alessio Colantoni [2], Gabriele Proietti[3,4], Jakob Rupert[1,2]

and Gian Gaetano Tartaglia [1,2,3,*]

[1] Center for Human Technology, Istituto Italiano di Tecnologia, Genoa,16152, Italy
[2] Department of Biology and Biotechnology Charles Darwin, Sapienza University of Rome, Rome, 00185, Italy
[3] Center for Life Nano Science, Istituto Italiano di Tecnologia, Rome, 00161, Italy
[4] Dipartimento di Neuroscienze, University of Genova, Genoa,16126, Italy

Present address: Alessio Colantoni, Center for Life Nano Science, Istituto Italiano di Tecnologia, Rome, 00161, Italy

* to whom correspondence should be addressed. Tel: +39 010 2897 621; Fax: +39 010 2897621; Email: gian.tartaglia@iit.it.

The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors.

MOTIF DATABASE CONSTRUCTION

The motif database was built based on information available in different databases and via literature mining. In addition to the species for which precompiled RBP and RNA libraries are available, the database also includes motifs from *Bos taurus, Oryctolagus cuniculus, Cricetulus griseus, Mesocricetus auratus, Gallus gallus, Xenopus laevis, Oryzias latipes, Tetraodon nigroviridis, Bombyx mori, Nematostella vectensis, Schistosoma mansoni, Arabidopsis thaliana, Zea mays, Physcomitrella patens, Thalassiosira pseudonana, Ostreococcus tauri, Neurospora crassa, Rhizopus oryzae, Vanderwaltozyma polyspora, Phytophthora ramorum, Leishmania major, Trypanosoma brucei, Trichomonas vaginalis, Plasmodium falciparum, Naegleria gruberi*. This allows to expand the number of motifs assigned by similarity.The databases we extracted motifs from are:

- **ATtRACT** (1). The database aggregates motif information from cisBP-RNA (2), RBPDB, SpliceAid-F (3) and Aedb (part of ASD (4)) databases with consensus RNA sequences obtained by analysing protein-RNA complexes deposited in the PDB database (5). A Position Probability Matrix (PPM) is available for each motif. Due to the heterogeneous nature of ATtRACT database, we applied some filters to improve the average motif quality. First, we discarded mutated RBPs and motifs obtained through techniques that we found to be extremely biased towards homopolymer sequences (*Homopolymer binding assay with recombinant protein, In vitro RNA-binding assay of Homopolymers and SDS-PAGE with recombinant protein, Homopolymer binding assay with HeLa cell/nuclear extracts,*

Immunoblotting of proteins selected by affinity chromatography with ribonucleotide homopolymer and HeLa nuclear extract, Fluorescence spectroscopy with recombinant protein, SDS-PAGE of ribonucleotide homopolymers with recombinant protein, Competition assay using homopolymers with HeLa S10 extracts, Competition assay using homopolymers with purified protein, Homopolymer binding assay and Western blot with HeLa extracts or recombinant protein, Immunoblots and Filter Binding Assay, Filter binding assay with purified protein, RNA affinity chromatography confirmed by UV crosslink and Western blot using HeLa nuclear extracts). Motifs were also filtered based on their length (minimum length of 5 nucleotides), quality score (0.01) and SpliceAid-F motif score (5).

- **cisBP-RNA** (2). We noticed that those cisBP-RNA motifs for which no PPM is available were not included in the ATtRACT database. IUPAC-encoded motifs were converted into PPMs and included in our motif database.

- **mCrossBase** (6). The database stores Position Frequency Matrices (PFMs) obtained by applying the mCross motif finding algorithm to ENCODE eCLIP data of 112 human RBPs. For each RBP, similar motifs are grouped into clusters; for each of these clusters, a representative motif is chosen based on the score. We extracted representative motifs with motif score > 40, converted them into PPMs and included them in our database.

- **oRNAment** (7). The motif database of this resource is composed of 218 PPMs obtained via RNAcompete, already available in cisBP-RNA database, and 235 PPMs identified by analysing RNA Bind-n-Seq (RBNS) data produced by the ENCODE project (8–10). The latter group of motifs was included in our database.

- **RBPmap** (11). This online resource includes a database of experimentally determined motifs.

We also compiled a literature-based dataset of motifs by extracting binding preferences (encoded as PPMs) from papers describing CLIP-Seq and similar experiments. To find such papers, we started from the list of experiments used to build ENCORI (formerly starBase v2.0 (12), available at http://starbase.sysu.edu.cn) and POSTAR2 (13) databases and we did further research to integrate this list with recent publications. Both ENCORI and POSTAR2 propose, for each RBP, one or more de novo motifs obtained by reanalysing the raw sequencing data. For each ENCORI *de novo* motif, the percentage of target and background sites containing the motif and a p-value are provided. High scoring ENCORI motifs were included in our database.

At the end of the motif collection phase, our database consisted of 539 PPMs in TRANSFAC format (14). For each PPM, we defined a core motif by trimming leading and trailing positions where maximum nucleotide probability was under 0.3. PPMs having a core motif length < 3 were discarded.

To avoid redundancy, RBPs with multiple motifs underwent a motif clustering procedure. First, we combined PPMs having the same consensus sequence into Familial Binding Profiles (FBP) (15) using STAMP software (16). An FBP is an “average” PPM of all the motifs used to build it. When such initial clustering did not result in a single motif, we performed another round of clustering using the STAMP

algorithm with parameters *-cc PCC -align SWU -forwardonly -chp -printpairwise -ma IR*. In case of two input motifs, we simply evaluated STAMP's motif alignment p-value: if it was < 0.1, the two motifs were considered as similar to each other and combined into an FBP; otherwise, they were kept as separate motifs. In case of more than two motifs, we accepted the clustering proposed by STAMP, unless it consisted in a single group of two motifs, with the remaining motifs organized as singletons. Such minimal clusterings were kept only when the following two conditions were both true:

- the two motifs of the largest cluster were similar. In case they were not, we opted for no clustering of motifs.
- no motif was similar to all the other motifs. In case we found such motifs, we combined all the motifs in a final FBP.

Instead of selecting representative motifs, we calculated FBPs and used them to represent clusters of motifs. PPMs were finally stripped to core motifs and converted to MEME format (17). Motifs were divided in four different categories depending on their length (3-, 4-, 5- and more than 5-letter motifs) (Figure SD1).

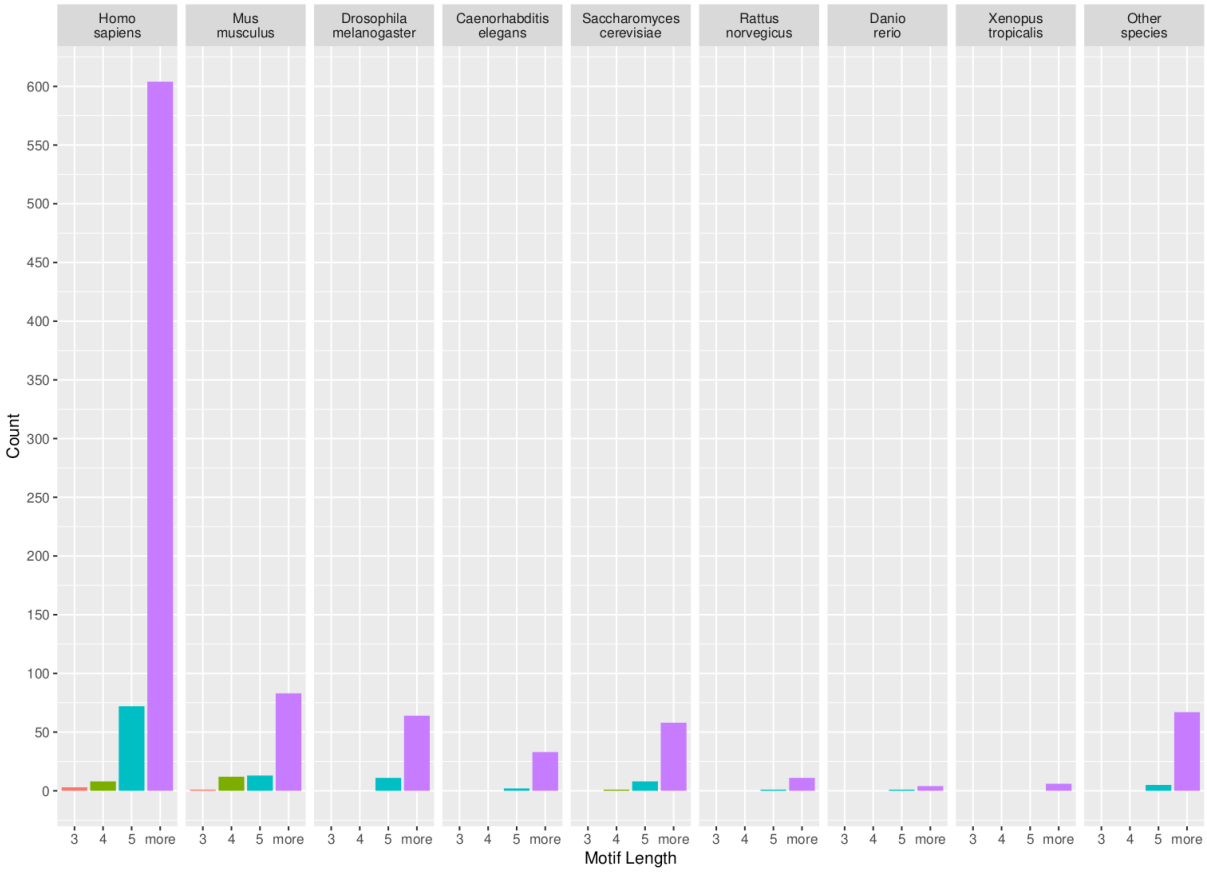

Figure SD1. Number of RNA binding motifs of different length for each organism. “Other species” refers to all the organisms that are not available in *cat*RAPID *omics v2.0* calculations, but which were used in the motif assignment protocol.

IDENTIFICATION OF MOTIF OCCURRENCES WITHIN RNA SEQUENCES

RNA sequences are scanned for individual matches to each of the motifs using FIMO software (18). Depending on the length of the motif, distinct FIMO p-value cutoffs are used (0.05, 0.01, 0.001, and 1e-4 for the 3-, 4-, 5- and more than 5-letter motifs, respectively). Occurrences of 3-letter and 4-letter motifs in a given RNA sequence are kept only if multiple instances are found occupying at least 8 and 7 positions in a window of 12, respectively; this ensures to have at least 3 close instances of a 3-letter motif or 2 close instances of a 4-letter motif without too much overlap between the motif occurrences.

REFERENCES FOR SUPPLEMENTARY METHODS

1. Giudice,G., Sánchez-Cabo,F., Torroja,C. and Lara-Pezzi,E. (2016) ATtRACT-a database of RNA-binding proteins and associated motifs. *Database*, **2016**.
2. Ray,D., Kazan,H., Cook,K.B., Weirauch,M.T., Najafabadi,H.S., 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.
3. Giulietti,M., Piva,F., D'Antonio,M., De Meo,P.D.O., Paoletti,D., Castrignanò,T., D'Erchia,A.M., Picardi,E., Zambelli,F., Principato,G., *et al.* (2013) SpliceAid-F: A database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res.*, **41**.
4. Stamm,S., Riethoven,J.J., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Tang,Y., Barbosa-Morais,N.L. and Thanaraj,T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, **34**, D46.
5. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
6. Feng,H., Bao,S., Rahman,M.A., Weyn-Vanhentenryck,S.M., Khan,A., Wong,J., Shah,A., Flynn,E.D., Krainer,A.R. and Zhang,C. (2019) Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites. *Mol. Cell*, **74**, 1189-1204.e6.
7. Benoit Bouvrette,L.P., Bovaird,S., Blanchette,M. and Lécuyer,E. (2020) ORNAment: A database of putative RNA binding protein target sites in the transcriptomes of model species. *Nucleic Acids Res.*, **48**, D166–D173.
8. Lambert,N., Robertson,A., Jangi,M., McGeary,S., Sharp,P.A. and Burge,C.B. (2014) RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Mol. Cell*, **54**, 887–900.
9. Lambert,N.J., Robertson,A.D. and Burge,C.B. (2015) RNA Bind-n-Seq: Measuring the binding affinity landscape of RNA-binding proteins. In *Methods in Enzymology*. Academic Press Inc., Vol. 558, pp. 465–493.
10. Dominguez,D., Freese,P., Alexis,M.S., Su,A., Hochman,M., Palden,T., Bazile,C., Lambert,N.J., Van Nostrand,E.L., Pratt,G.A., *et al.* (2018) Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol. Cell*, **70**, 854-867.e9.
11. Paz,I., Kosti,I., Ares,M., Cline,M. and Mandel-Gutfreund,Y. (2014) RBPmap: A web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.*, **42**, W361.
12. Li,J.H., Liu,S., Zhou,H., Qu,L.H. and Yang,J.H. (2014) StarBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.*, **42**.
13. Zhu,Y., Xu,G., Yang,Y.T., Xu,Z., Chen,X., Shi,B., Xie,D., Lu,Z.J. and Wang,P. (2019) POSTAR2: Deciphering the post-Transcriptional regulatory logics. *Nucleic Acids Res.*, **47**, D203–D211.
14. Wingender,E., Dietze,P., Karas,H. and Knüppel,R. (1996) TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
15. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.

16. Mahony,S. and Benos,P. V. (2007) STAMP: A web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253.
17. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202.
18. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.