Supplementary Material

RAIN: RNA–protein Association and Interaction Networks

Alexander Junge[1,2*], Jan C. Refsgaard[3*], Christian Garde[1,4*†], Xiaoyong Pan[1,2,3], Alberto Santos[3], Ferhat Alkan[1,2], Christian Anthon[1,2], Christian von Mering[5], Christopher T. Workman[1,4], Lars Juhl Jensen[1,3‡], Jan Gorodkin[1,2‡]

[1]Center for non-coding RNA in Technology and Health, University of Copenhagen, Denmark
[2]Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Denmark
[3]Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark
[4]Center for Biological Sequence Analysis, Technical University of Denmark, Denmark
[5]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland

[*]These authors contributed equally to this work.
[†]Current affiliation: The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[‡]To whom correspondence should be addressed: `lars.juhl.jensen@cpr.ku.dk`, `gorodkin@rth.dk`

1 RAIN summary

RAIN supplements the protein–protein interactions in STRING (1) with a set of ncRNA associations. Evidence for these ncRNA–target interactions can be separated into the following categories:

1. *Curated knowledge*
2. *Experimentally supported* interactions
3. *Predicted* miRNA–target interactions
4. Interactions identified by *text mining*

We integrate several publicly available resources of ncRNA associations in RAIN. Each resource yields evidence belonging to one of the categories listed above. Resources from each category are integrated as separate evidence channels, and each evidence channel assigns a *confidence score* to interactions.

If a certain ncRNA–target interaction is supported by multiple evidence channels, confidence scores from each channel are automatically combined by STRING. This allows the user to see how each evidence channel supports a given interaction directly in the STRING network view.

2 Curated knowledge

The first source of information comprises molecular interactions found in human that are well established in the scientific literature and/or listed in expert curated databases. This curated knowledge was collected for nine classes of ncRNAs, namely microRNA (miRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), signal recognition particle RNA (SRP RNA), Small nuclear RNA (snRNA), Small nucleolar RNA (snoRNA), Vault RNA, Y RNA and Telomerase RNA.

2.1 miRNA–mRNA interactions in Croft et al. gold standard data set

Expert curated miRNA–mRNA interactions were extracted from the Supplementary Material of Croft et al. (2). miRNA names and mRNA names were mapped to miRBase v20 (3) and STRING v10 (1) identifiers, respectively. The dataset comprises 152 interactions and also serves as part of gold standard in all subsequent analyses.

2.2 Ribosomal RNA (rRNA) and small nulceolar RNA (snoRNA)

Interactions of human protein and RNA components of the large and small ribosomal subunit were extracted from Reactome 49 (4). Additionally, we extracted the experimentally verified snoRNA and scaRNA interactions with human rRNAs from (5).

2.3 Transfer RNA (tRNA)

Human tRNAs were retrieved from the Genomic tRNA database (6) and identifiers assigned based on their HGNC annotations (7). Corresponding aminoacyl-tRNA synthetases were identified in the Aminoacyl-tRNA synthetases database (8) and mapped to STRING v10 (1) identifiers.

2.4 Signal recognition particle RNA (SRP RNA)

SRPDB (9) lists the following proteins as interaction partners of the human signal recognition particle RNA (7SL RNA): SRP9, SRP14, SRP19, SRP54, SRP68 and SRP72.

2.5 Vault RNA

The major vault protein (MVP) as well as the minor vault proteins (PARP4 and TEP1) have been shown with strong evidence to interact with vault RNA 1, vault RNA 2 and vault RNA 3 in human (10, 11, 12).

2.6 Y RNA

Human Ro60 has been shown to interact with Y RNAs in, e.g., Hogg and Collins (13), Stein et al. (14) and Green et al. (15) and human La (SSB autoantigen) interacts with Y5 RNA Hogg and Collins (13).

2.7 Telomerase RNA component (TERC)

TelomDB lists the following proteins as interaction partners of the human telomerase RNA component: TERT, DKC1, TEP1, TNKS, TERF2IP and TRF2 (`http://rnp.uthscsa.edu/rnp/telomDB/telomDB.html`; Accessed: 2015-03-26).

2.8 Spliceosomal RNA

The spliceosome is composed of multiple nuclear ribonucleoproteins. The interactions between spliceosomal proteins and RNAs (U1, U2, U4, U4atac, U6, U6atac, U11, U12) are annotated using Gene Ontology (GO) terms and added to RAIN. Proteins interacting with U1 spliceosomal RNA have one of the following annotations: GO:0034473 (U1 snRNA 3'-end processing), GO:0030619 (U1 snRNA binding) or GO:0005685 (U1 snRNP). Similar GO terms exist for all other U RNAs. These GO terms were extracted from Uniprot (16).

2.9 Small nucleolar RNA

RNA-RNA interactions of small nucleolar RNA were retrieved from the Human snoRNAome database (5). Note that, from this database, we only retrieved snoRNA interactions with reported modification site for antisense element.

3 Organisms and Interactions statistics

The tables in this section give an overview about the counts, sources and types of interactions in RAIN. The number of ncRNA associations per species contributed by each evidence channel is shown in Table 1. Table 2 contains the species-specific count of miRNAs, ncRNAs (miRNAs excluded) and mRNAs (or proteins translated from them) for which at least one interaction is available in RAIN. In all tables, only interactions with a combined score ≥ 0.15 are counted. These interactions can be queried in the RAIN web interface, while a much larger superset is available for download including interactions with scores below 0.15.

Table 1: Number of ncRNA interactions per organism in RAIN contributed by each resource type: curated, experimentally validated, predicted and text mined interactions.

Organism	Curated	Experiments	Predictions	Text mining	Total
H. sapiens	867	7611	162490	22315	188386
M. musculus	0	2575	71830	4609	77774
R. norvegicus	0	17	19192	926	20025
S. cerevisiae	0	385	0	340	725
Total	867	10588	253512	28190	

Table 2: Organism-specific number of miRNAs, ncRNAs (miRNAs excluded) and proteins for which at least one interaction is available in RAIN.

Organism	miRNAs	ncRNAs (miRNAs excluded)	mRNAs/proteins	Total
H. sapiens	2559	1449	17996	22004
M. musculus	1867	76	13758	15701
R. norvegicus	711	3	8640	9354
S. cerevisiae	0	96	200	296
Total	5137	1624	40594	47355

Table 3: Window size w used to fit the sigmoid transfer function f for each resource.

Resource	w
Experiments (combined)	40
miRanda	1000
miRDB	75
PITA	500
STarMirDB	200
TargetScan	50
Predictions (combined)	75
Text mining	100

4 miRNA predictor coverage of Tarbase validation set

Each miRNA target predictor integrated in RAIN only reports a subset of the total miRNA-mRNA interactions in the validation set due to a predictor specific score threshold. In Table 4, we list the fraction of the validation set that was reported by the respective miRNA predictors.

Table 4: Intersection of the validation set and the miRNA-mRNA pairs reported by the respective prediction methods, with ratio of this intersection to the size of the validation set listed in parenthesis.

Resource	Intersection with positive set	Intersection with negative set	Intersection with total validation set
PITA	1315 (95%)	439 (95%)	1754 (95%)
miRanda	1060 (76%)	284 (62%)	1342 (73%)
TargetScan	628 (45%)	197 (43%)	825 (45%)
miRDB	576 (42%)	83 (18%)	657 (36%)
STarMirDB	164 (15%)	50 (13%)	214 (14%)

5 Example: Combining confidence scores with and without accounting for the prior probability

Assume a given interaction is supported by text mining, prediction and experimental evidence, each with a confidence score of 0.01 - equal to the prior. Confidence score integration without accounting for the prior probability would yield a combined confidence score of

$$1 - (0.99 \cdot 0.99 \cdot 0.99) = 0.029701,$$

while accounting for the prior yields a combined confidence score of

$$1 - (1 - 0.01)^{-2} \cdot (0.99 \cdot 0.99 \cdot 0.99) = 0.01,$$

equal to the prior probability contributed by each evidence channel.

Similarly, for a confidence score of 0.02 in three evidence channels, not accounting for the prior probability yields a combined score of

$$1 - (0.98 \cdot 0.98 \cdot 0.98) = 0.058808$$

while acounting for the prior yields

$$1 - (1 - 0.01)^{-2} \cdot (0.98 \cdot 0.98 \cdot 0.98) = 0.03969799.$$

References

[1] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth A, Santos, K. Tsafou, M. Kuhn, P. Bork, Jensen L. J., and C. von Mering. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, 43(Database issue):D447–D452, Jan 2015.

[2] L. Croft, D. Szklarczyk, L. J. Jensen, and J. Gorodkin. Multiple independent analyses reveal only transcription factors as an enriched functional class associated with microRNAs. *BMC Syst. Biol.*, 6:90, 2012.

[3] A. Kozomara and S. Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, 42(Database issue):D68–D73, Jan 2014.

[4] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39(Database issue):D691–D697, Jan 2011.

[5] H. Jorjani, S. Kehr, D. J. Jedlinski, R. Gumienny, J. Hertel, P. F. Stadler, M. Zavolan, and A. R. Gruber. An updated human snoRNAome. *Nucleic Acids Res.*, 44(11):5068–5082, Jun 2016.

[6] T. M. Lowe and S. R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, 25(5):955–964, Mar 1997.

[7] K. A. Gray, B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.*, 43(Database issue):D1079–D1085, Jan 2015.

[8] J. Cavarelli, G. Eriani, B. Rees, M. Ruff, M. Boeglin, A. Mitschler, F. Martin, J. Gangloff, J. C. Thierry, and D. Moras. The active site of yeast aspartyl-tRNA synthetase: structural and functional aspects of the aminoacylation reaction. *EMBO J.*, 13(2):327–337, Jan 1994.

[9] M. A. Rosenblad, N. Larsen, T. Samuelsson, and C. Zwieb. Kinship in the SRP RNA family. *RNA Biol*, 6(5): 508–516, 2009.

[10] V. A. Kickhoefer, A. G. Stephen, L. Harrington, M. O. Robinson, and L. H. Rome. Vaults and telomerase share a common subunit, TEP1. *J. Biol. Chem.*, 274(46):32712–32717, Nov 1999.

[11] V. A. Kickhoefer, Y. Liu, L. B. Kong, B. E. Snow, P. L. Stewart, L. Harrington, and L. H. Rome. The telomerase/vault-associated protein TEP1 is required for vault RNA stability and its association with the vault particle. *J. Cell Biol.*, 152(1):157–164, Jan 2001.

[12] A. van Zon, M. H. Mossink, M. Schoester, G. L. Scheffer, R. J. Scheper, P. Sonneveld, and E. A. Wiemer. Multiple human vault RNAs. expression and association with the vault complex. *J. Biol. Chem.*, 276(40):37715–37721, Oct 2001.

[13] J. R. Hogg and K. Collins. Human Y5 RNA specializes a Ro ribonucleoprotein for 5S ribosomal RNA quality control. *Genes Dev.*, 21(23):3067–3072, Dec 2007.

[14] A. J. Stein, G. Fuchs, C. Fu, S. L. Wolin, and K. M. Reinisch. Structural insights into RNA quality control: the Ro autoantigen binds misfolded RNAs via its central cavity. *Cell*, 121(4):529–539, May 2005.

[15] C. D. Green, K. S. Long, H. Shi, and S. L. Wolin. Binding of the 60-kDa Ro autoantigen to Y RNAs: evidence for recognition in the major groove of a conserved helix. *RNA*, 4(7):750–765, Jul 1998.

[16] R. Apweiler, A. Bateman, M. J. Martin, C. O'Donovan, M. Magrane, Y. Alam-Faruque, E. Alpi, R. Antunes, J. Arganiska, E. Barrera Casanova, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, W. Mun Chan, G. Chavali, E. Cibrian-Uhalte, A. Da Silva, M. De Giorgi, F. Fazzini, P. Gane, L. G. Castro, P. Garmiri, E. Hatton-Ellis, R. Hieta, R. Huntley, D. Legge, W. Liu, J. Luo, A. MacDougall, P. Mutowo, A. Nightingale, S. Orchard, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, T. Sawford, A. Shypitsyna, E. Turner, V. Volynkin, T. Wardell, X. Watkins, H. Zellner, M. Corbett, M. Donnelly, P. van Rensburg, M. Goujon, H. McWilliam, R. Lopez, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, P. A. Binz, M. C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. de Castro, L. Cerutti, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, J. James, F. Jungo, G. Keller, V. Lara, P. Lemercier, J. Lew, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, B. E. Suzek, C. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, M. S. Yerramalla, and J. Zhang. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 42(Database issue):D191–D198, Jan 2014.