Published online 24 November 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D115–D118***
doi: 10.1093/nar/gkw1052

RAID v2.0: an updated resource of RNA-associated interactions across organisms

Ying Yi[1,†], Yue Zhao[1,2,†], Chunhua Li[1,†], Lin Zhang[1], Huiying Huang[1], Yana Li[1], Lanlan Liu[1], Ping Hou[1], Tianyu Cui[1,3], Puwen Tan[1], Yongfei Hu[1], Ting Zhang[1], Yan Huang[1], Xiaobo Li[2,*], Jia Yu[4,*] and Dong Wang[1,3,*]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China, [2]Department of Pathology, Harbin Medical University, Harbin 150081, China, [3]Department of Biochemistry and Molecular Biology, Shantou University Medical College, Shantou 515041, China and [4]State Key Laboratory of Medical Molecular Biology, Department of Biochemistry and Molecular Biology, School of Basic Sciences & Institute of Basic Medical Sciences, Peking Union Medical College & Chinese Academy of Medical Sciences, Beijing 100730, China

Received August 14, 2016; Revised October 18, 2016; Editorial Decision October 19, 2016; Accepted October 20, 2016

ABSTRACT

With the development of biotechnologies and computational prediction algorithms, the number of experimental and computational prediction RNA-associated interactions has grown rapidly in recent years. However, diverse RNA-associated interactions are scattered over a wide variety of resources and organisms, whereas a fully comprehensive view of diverse RNA-associated interactions is still not available for any species. Hence, we have updated the RAID database to version 2.0 (RAID v2.0, www.rna-society.org/raid/) by integrating experimental and computational prediction interactions from manually reading literature and other database resources under one common framework. The new developments in RAID v2.0 include (i) over 850-fold RNA-associated interactions, an enhancement compared to the previous version; (ii) numerous resources integrated with experimental or computational prediction evidence for each RNA-associated interaction; (iii) a reliability assessment for each RNA-associated interaction based on an integrative confidence score; and (iv) an increase of species coverage to 60. Consequently, RAID v2.0 recruits more than 5.27 million RNA-associated interactions, including more than 4 million RNA–RNA interactions and more than 1.2 million RNA–protein interactions, referring to nearly 130 000 RNA/protein symbols across 60 species.

INTRODUCTION

Recent developments have indicated that diverse RNA-associated (RNA–RNA/RNA–Protein) interactions are also fundamental to cellular processes like protein–protein interactions. They are also essential for a system-level understanding of cellular behavior (1–4). Hence, in recent years, a wide variety of experimental and computational prediction techniques have expanded a number of diverse RNA-associated interaction data sets. Most of these interactions are available in a variety of databases (5–9), including several databases that primarily manually collect and curate diverse RNA-associated interactions with experimental evidence from literature. Other databases focus on a more generalized perspective for diverse RNAs and their partners in specific cellular processes. Another resource predicts diverse RNA-associated interactions using computational prediction algorithms. However, a fully comprehensive view of diverse RNA-associated interactions is still not available for any particular species.

Because the comprehensive regulation of crosstalk between diverse RNA and proteins still remains ambiguous, we updated the RAID database (5) to version 2.0 (RAID v2.0, http://www.rna-society.org/raid/) by integrating experimental and computational prediction interactions through the manual curation of the literature and another 18 resources under one common framework (Figure 1). Accordingly, RAID v2.0 will offer several distinctive advantages: (i) integration from numerous resources, including experimental and computational prediction databases as well as manual curation of the literature (recruiting more than 5.27 million RNA-associated interactions and exceeding an 850-fold increase over the previous version); (ii) provision of an integrative confidence score for each RNA-associated in-

*To whom correspondence should be addressed. Tel: +86 451 86699584; Fax: +86 451 86699584; Email: wangdong@ems.hrbmu.edu.cn
Correspondence may also be addressed to Jia Yu. Tel: +86 10 69156423; Fax: +86 10 65240529; Email: j-yu@ibms.pumc.edu.cn
Correspondence may also be addressed to Xiaobo Li. Tel: +86 451 86699584; Fax: +86 451 86699584; Email: lixiaobo@ems.hrbmu.edu.cn
†These authors contributed equally to this work as the first authors.

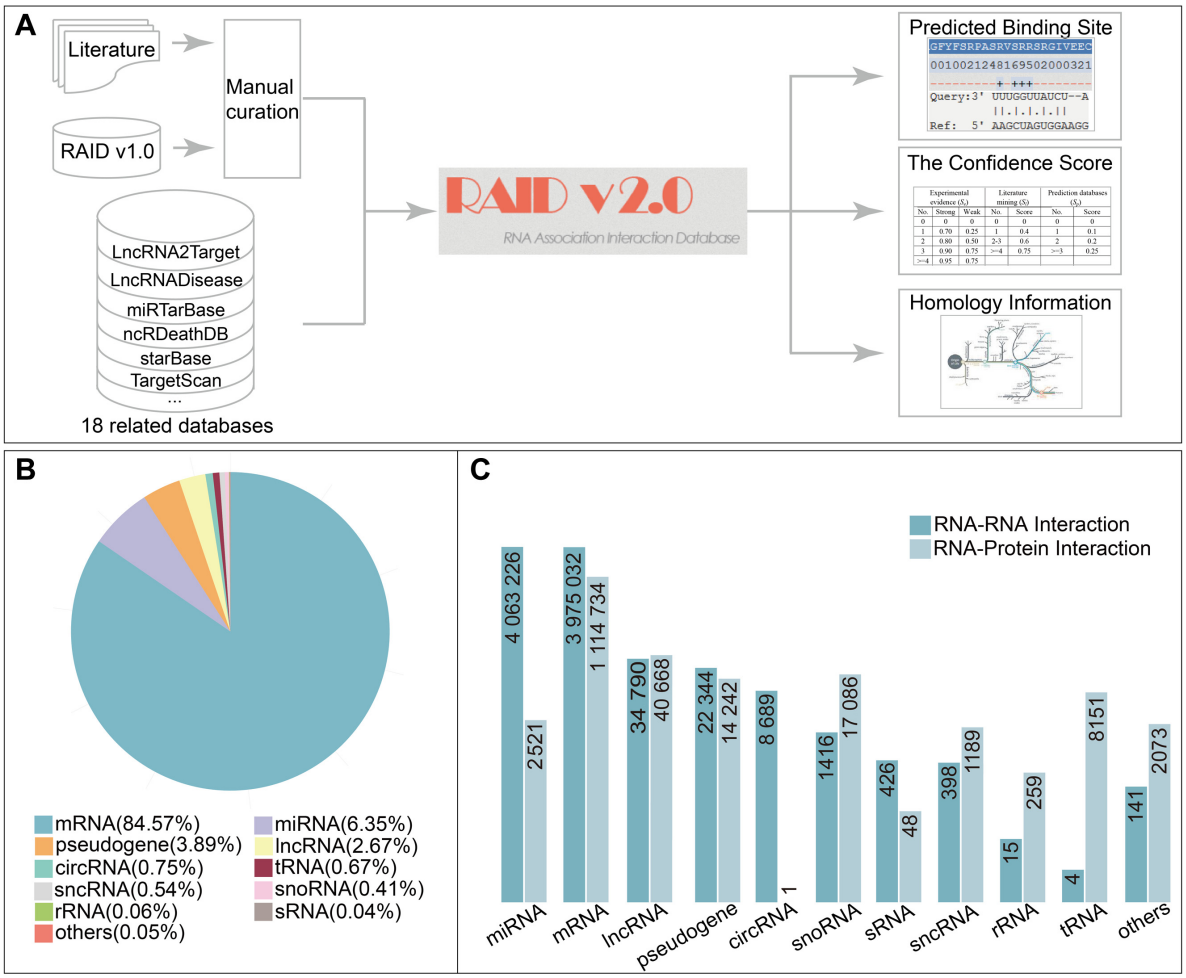

Figure 1. Flowchart of database construction and the statistics of RNA categories and interactions. (**A**) The overview of the RAID v2.0 database; (**B**) The percentage of diverse RNA categories in RAID v2.0 database; (**C**) The number of RNA–RNA/RNA–protein interactions for diverse RNA categories in RAID v2.0 database, the height of histogram transformed by log10.

teraction, considering that an integrated scoring strategy will offer higher confidence when independent types of evidence agree; and (iii) mapping RNA-associated interactions into numerous species to facilitate studies of homology (increased coverage across 60 species).

DATA COLLECTION

To update this version of the RAID database, we first screened all of the literature in the PubMed database (mainly from 2000–2016) with the following keywords combinations: (i) RNA–RNA interactions: (RNA symbols or RNA category names) and (RNA symbols or RNA category names) and (e.g. interaction or binding); (ii) RNA–protein interactions: (RNA symbols or RNA category names) and (protein symbols) and (e.g. interaction or binding). The relevant hits were downloaded and prepared systematically for further manual data curation. Second, RAID v2.0 integrated diverse RNA-associated interactions from other 18 databases, including ChIPBase (10), LncRNA2Target (11), LncRNAdisease (7), miR2Disease (12), miRTarBase (13), MNDR (14), ncRDeathDB (8), NPInter (15), OncomiRDB (16), sRNATarBase (17), StarBase (6), TransmiR (18) and ViRBase (19) as well as five computational prediction databases (DroID (20), EIMMo (21), miRanda (22), miRDB (23) and TargetScan (9)).

For the RNA/protein names collected from different resources, RAID v2.0 mapped these symbols to either an official gene Symbol or a miRBase ID and presented them to NCBI Alias, HGNC ID, Ensembl ID, OMIM ID, HPRD ID and UniProtKB protein accession, among others. Furthermore, to facilitate researcher access to information from external resources, we also linked Entrez ID, miRBase accession and UniprotKB protein accession to the NCBI Gene, miRBase database and UniProt (24) database, which can efficiently retrieve a substantial amount of genomic-associated data from external resources.

INTEGRATIVE CONFIDENCE SCORES

In RAID v2.0, the RNA-associated interactions are collected from different types of resources under one common framework, including experimental, literature mining and computational prediction evidence. Furthermore, sim-

ilar to miRTarBase database, the experimental evidence in RAID v2.0 was divided into strong experimental evidence (e.g. RNA immunoprecipitation and luciferase reporter assay) and weak experimental evidence (e.g. ChIP-seq and CLIP-seq) by a manual assignment, depending on the nature and qualitative annotation of the experiment method. Because multiple types of evidence contribute to the identification of a specific RNA-associated interaction, the RNA-associated interactions stored in RAID v2.0 are not equally reliable. Because it is difficult for a user to assess the quality of each interaction, we developed an integrative confidence score system to facilitate the evaluation of the reliability of each RNA-associated interaction (25). An integrative confidence score that combines scores from all of these evidence resources can give an overall estimation of the reliability of each RNA-associated interaction.

In principle, we assume that (i) experimental evidence contributes more significantly to the confidence score than does evidence derived from computational prediction algorithms; (ii) strong experimental evidence with lower false positive rates are considered to provide more reliable evidence than weak experimental evidence; and (iii) RNA-associated interactions supported by more evidence resources should be given higher confidence scores than those supported by fewer evidence resources. Therefore, we firstly assign quantitative confidence scores (strong experimental evidence: s_s, weak experimental evidence: s_w, computational prediction database: s_p) to each RNA-associated interaction based on the evidence types and number of evidence resources as follows:

$$s_i = \begin{cases} 0, & x = 0 \\ \frac{w_i}{1+e^{-x}}, & x > 0 \end{cases} \quad (1)$$

where i is the evidence type (s_s: strong experimental evidence, s_w: weak experimental evidence, s_p: computational prediction database) and x is the number of evidence resources, we set weight factor w_s, w_w and w_p to 1, 0.75 and 0.25, respectively.

Finally, an integrative confidence score (S) is calculated as:

$$S = 1 - \prod_i (1 - s_i) \quad (2)$$

Hence, as illustrated in Supplementary Figure S1, our integrative confidence score system can effectively estimate the reliability of each RNA-associated interaction with more or fewer evidence types and the number of resources. The resulting score ranges between 0 and 1. Only well-supported interactions obtain a value close to 1. Therefore, this is an effective tool for filtering interactions.

DATABASE CONTENT AND CONSTRUCTION

In total, RAID v2.0 recruits 5,272,396 RNA-associated entries (an over 850-fold increase from the previous version), including over 4 million RNA–RNA interactions and over 1.2 million RNA–protein interactions, referring to 129 857 RNA/protein symbols. RAID v2.0 involves at least 13 RNAs (including circRNA, lncRNA, miRNA, mRNA, miscRNA, pseudogenes, rRNA, scRNA, sncRNA, snoRNA, snRNA, sRNA and tRNA) and contains up to 60 species covering seven categories (bacteria, fungi, insects, nematodes, plants, vertebrates and viruses). More importantly, each RNA-associated interaction in RAID v2.0 is provided with an integrative confidence score. The user can select RNA-associated interactions by a user-specific threshold.

A 'Homology' option has been added to the 'Detail Information' page to help users investigate the conservation of RNA-associated interactions between RNA orthology/paralogy obtained from miRBase and NCBI HomoloGene (Supplementary Figure S2). In the current version, there are more than 80 000 RNAs/proteins with homology information.

In RAID v2.0, we have also modified the display of the predicted binding sites for RNA-associated interactions because several tools used in the previous version were not available. For RNA–RNA interactions, the binding sites and scores are predicted according to miRanda (22) and RISearch (26). For RNA–protein interactions, PRIdictor is used to predict RNA-binding residues in proteins. Additionally, RAID v2.0 have represented the experimental verified RNA-binding sites in proteins documented in RBPDB (27), RsiteDB (28) and PDB (29) databases (Supplementary Figure S2).

On the updated 'Browse' page, users can access RAID v2.0 via three different paths: 'Interaction Type', 'Species' and 'Detection Methods'. For user convenience, we have designed the treeview and users can obtain browse results by clicking the node.

CONCLUSION AND FUTURE DIRECTIONS

In past decades, numerous protein–protein interactions databases have been established, including the most widely used STRING database. This has led to a more comprehensive understanding of protein functions and cellular processes. However, recent developments have indicated that protein–protein interactions represent perhaps only half of the story in cells. The RNA-associated interactome is likely to be much larger and more complex than we can imagine. Currently, diverse RNA-associated interactions are scattered over a wide variety of resources and organisms. A fully comprehensive view of all diverse RNA-associated interactions is still not available for any species. Consequently, we have updated the RAID database to version 2.0 by integrating manually reading literature and 18 other database resources under one common framework and providing an integrative confidence score for each RNA-associated interaction. RAID v2.0 aims to provide a comprehensive and reliably assessed collection of RNA-associated interactions across organisms. Furthermore, because each RNA-associated interaction has an integrative confidence score, users can filter the diverse RNA-associated interaction network at any threshold.

In the future, we will expand the database with more information, including RNA binding domain annotation, 2D and 3D RNA structures and improvement of the current computational prediction algorithm to obtain our own predicted data. With the emergence of more RNA-related information, we may improve the integrative confidence scoring strategy. We will keep a watchful eye on new research

progress and will continuously curate and update the reference data. Hence, complemented by the successful PPI databases, RAID will provide a valuable skeleton for better understanding the functional organization of the cell.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Natural Science Foundation of Heilongjiang Province of China [C2015027]; Scientific Research Fund of Heilongjiang Provincial Education Department [12541426]; WeihanYu Youth Science Fund Project of Harbin Medical University. Funding for open access charge: Natural Science Foundation of Heilongjiang Province of China [C2015027]; Scientific Research Fund of Heilongjiang Provincial Education Department [12541426]; WeihanYu Youth Science Fund Project of Harbin Medical University.
Conflict of interest statement. None declared.

REFERENCES

1. Sumazin,P., Yang,X., Chiu,H.S., Chung,W.J., Iyer,A., Llobet-Navas,D., Rajbhandari,P., Bansal,M., Guarnieri,P., Silva,J. *et al.* (2011) An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell*, **147**, 370–381.
2. Guttman,M. and Rinn,J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346.
3. Huerta-Cepas,J., Szklarczyk,D., Forslund,K., Cook,H., Heller,D., Walter,M.C., Rattei,T., Mende,D.R., Sunagawa,S., Kuhn,M. *et al.* (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.*, **44**, D286–D293.
4. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.
5. Zhang,X., Wu,D., Chen,L., Li,X., Yang,J., Fan,D., Dong,T., Liu,M., Tan,P., Xu,J. *et al.* (2014) RAID: a comprehensive resource for human RNA-associated (RNA-RNA/RNA-protein) interaction. *RNA*, **20**, 989–993.
6. 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**, D92–D97.
7. Chen,G., Wang,Z., Wang,D., Qiu,C., Liu,M., Chen,X., Zhang,Q., Yan,G. and Cui,Q. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.
8. Wu,D., Huang,Y., Kang,J., Li,K., Bi,X., Zhang,T., Jin,N., Hu,Y., Tan,P., Zhang,L. *et al.* (2015) ncRDeathDB: A comprehensive bioinformatics resource for deciphering network organization of the ncRNA-mediated cell death system. *Autophagy*, **11**, 1917–1926.
9. Agarwal,V., Bell,G.W., Nam,J.W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.
10. Yang,J.H., Li,J.H., Jiang,S., Zhou,H. and Qu,L.H. (2013) ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. *Nucleic Acids Res.*, **41**, D177–D187.
11. Jiang,Q., Wang,J., Wu,X., Ma,R., Zhang,T., Jin,S., Han,Z., Tan,R., Peng,J., Liu,G. *et al.* (2015) LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression. *Nucleic Acids Res.*, **43**, D193–D196.
12. Jiang,Q., Wang,Y., Hao,Y., Juan,L., Teng,M., Zhang,X., Li,M., Wang,G. and Liu,Y. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**, D98–D104.
13. Chou,C.H., Chang,N.W., Shrestha,S., Hsu,S.D., Lin,Y.L., Lee,W.H., Yang,C.D., Hong,H.C., Wei,T.Y., Tu,S.J. *et al.* (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*, **44**, D239–D247.
14. Wang,Y., Chen,L., Chen,B., Li,X., Kang,J., Fan,K., Hu,Y., Xu,J., Yi,L., Yang,J. *et al.* (2013) Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network. *Cell Death Dis.*, **4**, e765.
15. Hao,Y., Wu,W., Li,H., Yuan,J., Luo,J., Zhao,Y. and Chen,R. (2016) NPInter v3.0: an upgraded database of noncoding RNA-associated interactions. *Database (Oxford)*, **2016**, baw057.
16. Wang,D., Gu,J., Wang,T. and Ding,Z. (2014) OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs. *Bioinformatics*, **30**, 2237–2238.
17. Wang,J., Liu,T., Zhao,B., Lu,Q., Wang,Z., Cao,Y. and Li,W. (2016) sRNATarBase 3.0: an updated database for sRNA-target interactions in bacteria. *Nucleic Acids Res.*, **44**, D248–D253.
18. Wang,J., Lu,M., Qiu,C. and Cui,Q. (2010) TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Res.*, **38**, D119–D122.
19. Li,Y., Wang,C., Miao,Z., Bi,X., Wu,D., Jin,N., Wang,L., Wu,H., Qian,K., Li,C. *et al.* (2015) ViRBase: a resource for virus-host ncRNA-associated interactions. *Nucleic Acids Res.*, **43**, D578–D582.
20. Murali,T., Pacifico,S., Yu,J., Guest,S., Roberts,G.G. 3rd and Finley,R.L. Jr (2011) DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila. *Nucleic Acids Res.*, **39**, D736–D743.
21. Gaidatzis,D., van Nimwegen,E., Hausser,J. and Zavolan,M. (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 69.
22. Betel,D., Koppal,A., Agius,P., Sander,C. and Leslie,C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, **11**, R90.
23. Wong,N. and Wang,X. (2015) miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res.*, **43**, D146–D152.
24. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
25. Guo,J., Liu,H. and Zheng,J. (2016) SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic Acids Res.*, **44**, D1011–D1017.
26. Wenzel,A., Akbasli,E. and Gorodkin,J. (2012) RIsearch: fast RNA-RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics*, **28**, 2738–2746.
27. Cook,K.B., Kazan,H., Zuberi,K., Morris,Q. and Hughes,T.R. (2011) RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.*, **39**, D301–D308.
28. Shulman-Peleg,A., Nussinov,R. and Wolfson,H.J. (2009) RsiteDB: a database of protein binding pockets that interact with RNA nucleotide bases. *Nucleic Acids Res.*, **37**, D369–D373.
29. Rose,P.W., Prlic,A., Bi,C., Bluhm,W.F., Christie,C.H., Dutta,S., Green,R.K., Goodsell,D.S., Westbrook,J.D., Woo,J. *et al.* (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, **43**, D345–D356.