doi: 10.1093/nar/gkw728

RNALocate: a resource for RNA subcellular localizations

Ting Zhang[1,†], Puwen Tan[1,†], Liqiang Wang[1,†], Nana Jin[1,†], Yana Li[1], Lin Zhang[1], Huan Yang[2], Zhenyu Hu[1], Lining Zhang[1], Chunyu Hu[1], Chunhua Li[1], Kun Qian[1], Changjian Zhang[2], Yan Huang[1], Kongning Li[1,*], Hao Lin[2,*] and Dong Wang[1,3,*]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China, [2]Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China and [3]Department of Biochemistry and Molecular Biology, Shantou University Medical College, Shantou 515041, China

Received July 12, 2016; Accepted August 08, 2016

ABSTRACT

Increasing evidence has revealed that RNA subcellular localization is a very important feature for deeply understanding RNA's biological functions after being transported into intra- or extra-cellular regions. RNALocate is a web-accessible database that aims to provide a high-quality RNA subcellular localization resource and facilitate future researches on RNA function or structure. The current version of RNALocate documents more than 37 700 manually curated RNA subcellular localization entries with experimental evidence, involving more than 21 800 RNAs with 42 subcellular localizations in 65 species, mainly including *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae* etc. Besides, RNA homology, sequence and interaction data have also been integrated into RNALocate. Users can access these data through online search, browse, blast and visualization tools. In conclusion, RNALocate will be of help in elucidating the entirety of RNA subcellular localization, and developing new prediction methods. The database is available at http://www.rna-society.org/rnalocate/.

INTRODUCTION

Biological functions of RNAs, including translation of genetic information, cellular signal transduction and transcriptional regulation etc., are determined by their location in cell (1,2). A cell is divided into different compartments that are related to different biological processes (3). For example, the RNA localized in nuclear usually participates in gene expression regulation or mitosis etc (4). Thus, the cellular role of the RNA after synthesis could be inferred from its subcellular localization information. Based on this, subcellular localization for RNAs plays an important role in studying biological function of RNAs. Therefore, it is urgent to construct a database of RNA subcellular localization to integrate, analyze and predict RNA subcellular localization for speeding up RNA structural and functional research.

To complement with related research in RNA subcellular localization, we developed a web-accessible database (RNALocate, http://www.rna-society.org/rnalocate/), aimed to collect expanding catalog of diverse species' RNA subcellular localization in multiple biological processes by manually curating the literature. The first release of RNALocate has contained more than 37 700 manually curated RNA subcellular localization entries with experimental evidence, involving 65 organisms (such as *Homo sapiens, Musmusculus and Saccharomyces cerevisiae*), 42 subcellular localizations (such as Cytoplasm, Nucleus, Endoplasmic reticulum, Ribosome) and 9 RNA categories (such as mRNA, miRNA, lncRNA). Hence, RNALocate provides a more specific subcellular localization resource in which to efficiently investigate, browse and analyze a particular RNA, and even provides insight into the functions of hypothetical or novel RNAs. The whole data set can be easily queried and downloaded through the webpage, and visualization tools for interactively browsing and analyzing the data set are provided. In addition, RNALocate also allows researchers to submit new RNA subcellular localization.

[*]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 Hao Lin. Tel: +86 28 83202351; Fax: +86 28 83208238; Email: hlin@uestc.edu.cn
Correspondence may also be addressed to Kongning Li. Tel: +86 451 86615922; Fax: +86 451 86615922; Email: kongningli@hotmail.com
[†]These authors contributed equally to the work as first authors.

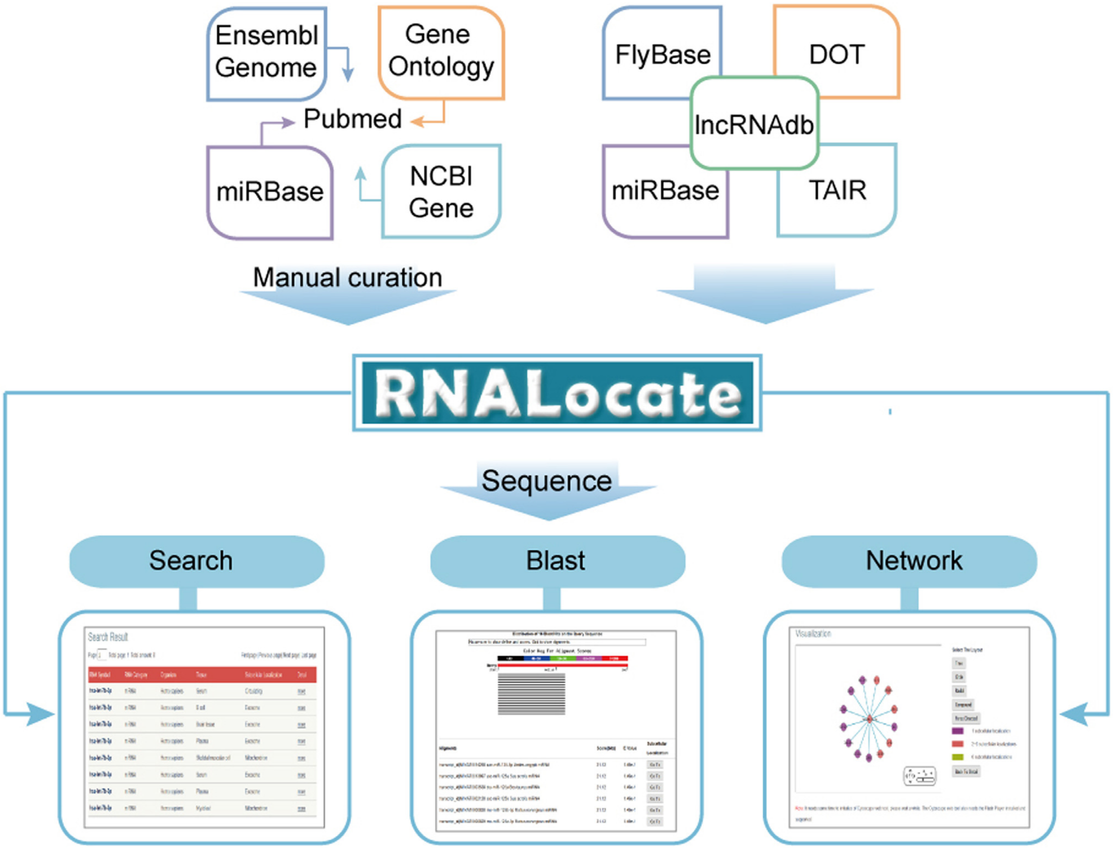

Figure 1. The overview of the RNALocate database.

DATA SOURCES AND IMPLEMENTATION

In order to collect all available RNAs, RNALocate integrates all types of RNA symbols, mainly including microRNA symbols from the miRBase database (5), long noncoding RNA (lncRNA) and mRNA symbols from NCBI Gene and Ensemble genome database (6,7). Other ncRNA category names are also included, such as transfer RNA and small nuclear RNA from NCBI Gene and Ensemble genome database (6,7). The list of subcellular localization names was collected according to the Gene Ontology (GO) (8). We have written a simple script to screen all abstracts and articles in the PubMed database using the following keyword combinations: (each RNA symbol or RNA category name) and/or (each subcellular localization). The relevant hits were further inspected manually. Besides, we also retrieved several thousand subcellular localization entries from lncRNAdb (9), PomBase (10), FlyBase (11), TAIR (12) and DOT (13) databases (Figure 1).

The RNALocate database is implemented using HTML and PHP languages with MySQL server. The interface component consists of web pages designed and implemented in HTML/CSS. It has been tested in the Google Chrome, Firefox and Internet Explorer web browsers.

DATABASE CONTENT

RNA subcellular localization information was manually obtained from articles published in the PubMed database before May 2016. In current version, RNALocate documents 37 772 RNA subcellular localization entries with experimental evidence from 65 organisms, involving 42 subcellular localizations (Figure 2) and 9 RNA categories (including csRNA, lncRNA, mRNA, miRNA, piRNA, snRNA, rRNA, snoRNA and tRNA) (Figure 3). Among these, more than 1400 entries were collected from lncRNAdb, PomBase, FlyBase, TAIR and DOT databases. Each subcellular localization entry contains detailed information on RNA symbol, RNA category, alias, organism, sequence, homology, subcellular localization, tissue, validation method, PubMed ID, detailed description and network.

In 'Submit' page, RNALocate invites users to upload novel RNA subcellular localization data, and in 'Blast' page, sequence alignment can be done after parameter selection. Except these, the whole data set could be downloaded through two approaches: 'Basic Download' and 'API' (application programming interface). In 'Basic Download' page, the whole data are saved in Microsoft Excel and TXT formats, users can get them by clicking the download button. In 'API' page, users can access part of RNALocate data by using script. RNALocate also provides three options in 'Help' page to supply instructions for using it, including 'Statistics' (detailed statistical tables), 'Tutorial' (procedure and illustrations of RNALocate) and 'Sister Databases'.

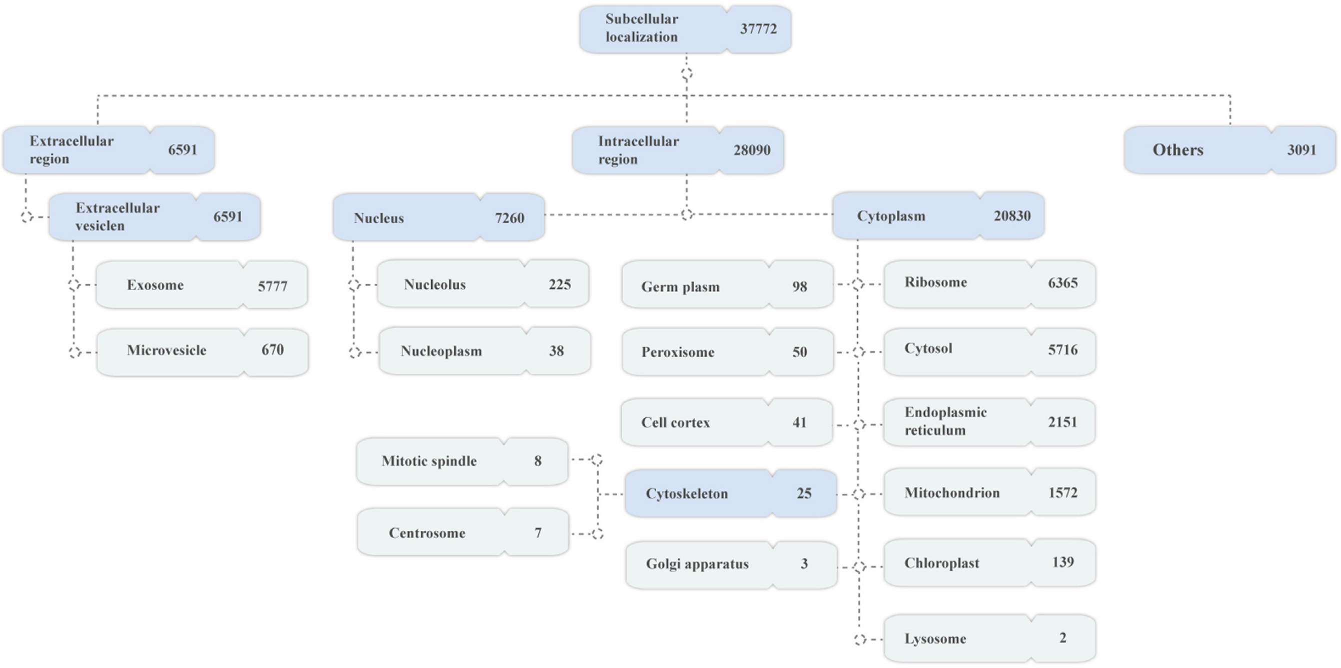

Figure 2. The hierarchical organization and statistics of RNA subcellular localization.

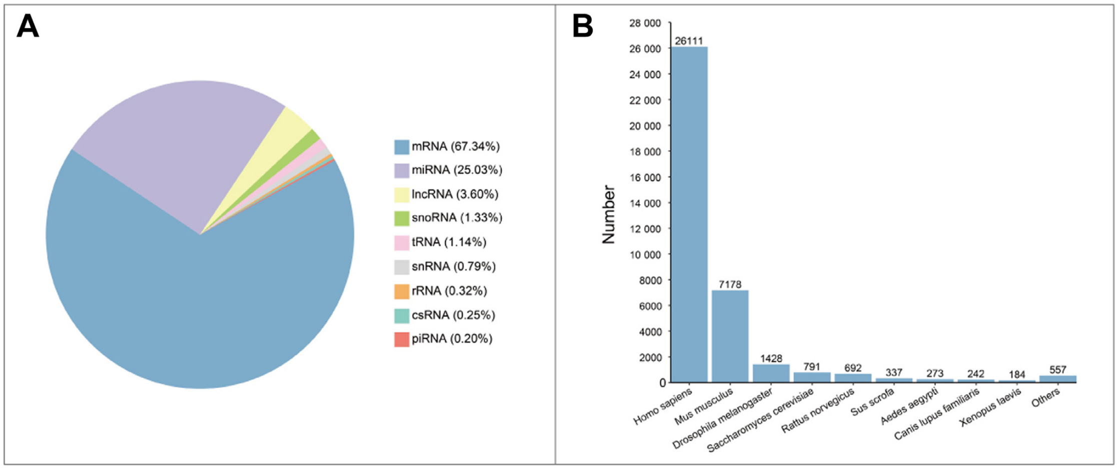

Figure 3. The statistics of RNA category and organism. (**A**) The percentage of 9 RNA categories in RNALocate database (**B**) The entry number of 65 organism in RNALocate database, only the organisms with ≥100 entries have been listed, respectively.

DATA QUERYING, SEARCHING AND BROWSING

RNALocate provides an interface for convenient retrieval of all RNA subcellular localizations. Users can query each entry through ‘Keyword Search’ in ‘Search’ page. In ‘Keyword Search’, 5 paths and relevant examples have been provided, including ‘RNA Symbol’, ‘RNA Category’, ‘Subcelllular localization’, ‘Organism’ and ‘Other ID (miRBase ID/Entrez ID)’.

RNALocate provides brief details of search results as a table in the ‘Search Result’ page, while more detailed descriptions such as PubMed ID and description of the reference are displayed in ‘Detail’ page reached by selecting ‘more’. When selecting the specific RNA symbol in ‘Search Result’ page, the ‘Detail’ page presents more associated information of the RNA, such as organism, subcellular localization, alias, sequence, homology and validated method. More than 9200 RNAs with orthology/paralogy from miRBase and Homologene database have been provided in RNALocate for investigation on RNA subcellular localizations conservation. To further understand the interaction information between different RNAs in various types of subcellular localizations online, a ‘Network’ option has also been provided to visualize RNA interaction network with subcellular localization and organism (14,15).

In ‘Browse’ page, users can access RNALocate in three different paths: ‘By Localization’, ‘By RNA Category’ and ‘By Organism’. A treeview and figure have been displayed in the three pages, respectively. Users could get browse results by clicking the node on the tree or the associated name in the figure. For convenience, the data in RNALocate are organized using a hierarchical structure of subcellular localization, according to the cellular component annotations documented in GO (8).

DISCUSSION AND FUTURE PROSPECTS

Several subcellular localization databases focused on proteins have been previously constructed, such as DBSubLoc, Organelle DB, eSLDB, LOCATE, SUBA, LocDB and PSORTdb databases (3,16–21). They had led to a more comprehensive understanding of the biological functions in proteins. However, recent development has indicated that protein subcellular localization are perhaps only half of the story in cells, since an expanding catalog of diverse RNAs is actively involved in multiple biological processes in different subcellular localization. To complement with this absence, we developed the RNALocate database by organizing and presenting RNA subcellular localization data for 65 organisms across 9 RNA categories. To our knowledge, this is the first database comprehensively focusing on RNA subcellular localization. We hope this resource will bridge the gap in RNAs and subcellular localization research, and stimulate further interesting elucidating the entirety of RNA subcellular localization, and developing new prediction methods. In the future, we will continuously collate RNA subcellular localization reference data and update RNALocate.

FUNDING

Scientific Research Fund of Heilongjiang Provincial Education Department [12541426].

Conflict of interest statement. None declared.

REFERENCES

1. Martin,K.C. and Ephrussi,A. (2009) mRNA localization: gene expression in the spatial dimension. *Cell*, **136**, 719–730.
2. Kuersten,S. and Goodwin,E.B. (2003) The power of the 3′ UTR: translational control and development. *Nat. Rev. Genet.*, **4**, 626–637.
3. Sprenger,J., Lynn Fink,J., Karunaratne,S., Hanson,K., Hamilton,N.A. and Teasdale,R.D. (2008) LOCATE: a mammalian protein subcellular localization database. *Nucleic Acids Res.*, **36**, D230–D233.
4. van Heesch,S., van Iterson,M., Jacobi,J., Boymans,S., Essers,P.B., de Bruijn,E., Hao,W., MacInnes,A.W., Cuppen,E. and Simonis,M. (2014) Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol.*, **15**, R6.
5. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
6. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
7. Stalker,J., Gibbins,B., Meidl,P., Smith,J., Spooner,W., Hotz,H.R. and Cox,A.V. (2004) The Ensembl Web site: Mechanics of a genome browser. *Genome Res.*, **14**, 951–955.
8. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
9. Quek,X.C., Thomson,D.W., Maag,J.L., Bartonicek,N., Signal,B., Clark,M.B., Gloss,B.S. and Dinger,M.E. (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.*, **43**, D168–D173.
10. McDowall,M.D., Harris,M.A., Lock,A., Rutherford,K., Staines,D.M., Bahler,J., Kersey,P.J., Oliver,S.G. and Wood,V. (2015) PomBase 2015: updates to the fission yeast database. *Nucleic Acids Res.*, **43**, D656–D661.
11. Drysdale,R. (2008) FlyBase : A database for the Drosophila research community. *Methods Mol. Biol.*, **420**, 45–59.
12. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
13. Jambor,H., Surendranath,V., Kalinka,A.T., Mejstrik,P., Saalfeld,S. and Tomancak,P. (2015) Systematic imaging reveals features and changing localization of mRNAs in Drosophila development. *eLife*, **4**, e05003.
14. 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.
15. 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.
16. Guo,T., Hua,S., Ji,X. and Sun,Z. (2004) DBSubLoc: database of protein subcellular localization. *Nucleic Acids Res.*, **32**, D122–D124.
17. Peabody,M.A., Laird,M.R., Vlasschaert,C., Lo,R. and Brinkman,F.S. (2016) PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. *Nucleic Acids Res.*, **44**, D663–D668.
18. Pierleoni,A., Martelli,P.L., Fariselli,P. and Casadio,R. (2007) eSLDB: Eukaryotic subcellular localization database. *Nucleic Acids Res.*, **35**, D208–D212.
19. Rastogi,S. and Rost,B. (2011) LocDB: experimental annotations of localization for Homo sapiens and Arabidopsis thaliana. *Nucleic Acids Res.*, **39**, D230–D234.
20. Tanz,S.K., Castleden,I., Hooper,C.M., Vacher,M., Small,I. and Millar,H.A. (2013) SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in Arabidopsis. *Nucleic Acids Res.*, **41**, D1185–D1191.
21. Wiwatwattana,N., Landau,C.M., Cope,G.J., Harp,G.A. and Kumar,A. (2007) Organelle DB: an updated resource of eukaryotic protein localization and function. *Nucleic Acids Res.*, **35**, D810–D814.