Data in Brief 14 (2017) 143–147

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Consensus datasets of mouse miRNA-mRNA interactions from multiple online resources

Zhao Bi, Bin Xue*

Department of Cell Biology, Microbiology and Molecular Biology, School of Natural Sciences and Mathematics, College of Arts and Sciences, University of South Florida, 4202 East Fowler Ave. ISA2015, Tampa, FL 33620, USA

ARTICLE INFO

Article history:
Received 20 April 2017
Received in revised form
19 June 2017
Accepted 18 July 2017
Available online 20 July 2017

ABSTRACT

MiRNAs regulate gene expression by forming base pairing with mRNAs to inhibit the translation of those mRNAs. In many mammalian genomes each, about 2000 miRNAs were found to regulate roughly 60% of all the genes in that genome. Many experimental validations and computational predictions have been done on miRNA:mRNA interactions. Nonetheless, the interactions from different sources are not always consistent. In this study, we integrated multiple online resources, including mirTarBase, TarBase, miRanda, miRDB, PITA, and TargetScan, and developed eleven large-scale datasets containing miRNA:mRNA interactions that are consistent among a specific subgroup of above-mentioned online resources. In addition, a new integrated confidence score was designed to show the significance for all the miRNA:mRNA interactions.

Specifications Table

Subject area	*Biology*
More specific subject area	*Genomics, non-coding RNA, miRNA, gene expression*
Type of data	*Excel file*

DOI of original article: http://dx.doi.org/10.1016/j.ygeno.2017.04.003

* Corresponding author.
E-mail address: binxue@usf.edu (B. Xue).

http://dx.doi.org/10.1016/j.dib.2017.07.035
2352-3409/

How data was acquired	*Databases*
Data format	*Raw, filtered, analyzed*
Experimental factors	*Experimentally validated data were incorporated*
Experimental features	*Consistent data were selected from multiple sources*
Data source location	*Online resources*
Data accessibility	*Within this article*

Value of the data

- The eleven datasets are the largest consistent datasets of miRNA:mRNA interactions containing both true and false samples in mouse genome.
- These datasets can be further refined as training and test datasets in the development of new miRNA target predictors, and therefore are critical for bioinformatics studies.
- The samples in these datasets are genome-wide and are consistent among multiple sources. Therefore, these datasets can be used directly to extract high-confidence miRNA:mRNA interactions.

1. Data

MiRNAs are short non-coding RNA molecules with about 22 nucleotides. Although being very short compared to mRNAs, which normally have hundreds to thousands of nucleotides, miRNAs may interact with the 3′-UTRs of mRNAs to inhibit the translation of those mRNAs. MiRNAs are abundant in natural, there are more than 2000 miRNAs in each mammalian genome, regulating about 60% of all the genes in that genome. Clearly, miRNAs are a critical family of gene expression regulators. However, a thorough understanding on the function and mechanism of miRNAs is still elusive. To facilitate the studies of miRNA molecular biology, curated datasets of miRNA:mRNA interactions will be very helpful. The eleven datasets developed in this project are the largest consistent datasets containing both true and false miRNA:mRNA interactions in mouse genome. An integrated confidence score has also been designed to show the significance of these interactions.

2. Experimental design, materials and methods

The consistent datasets were developed using the following online resources: miRNA:mRNA interaction data of mouse genome in mirTarBase Release 6.0 [1] and TarBase v7.0 [2], pre-assembled datasets for predicted miRNA:mRNA interactions in mouse genome downloaded from the web servers of miRanda 2010 release [3], miRDB v5.0 [4], PITA v6 [5], and TargetScan 7.0 [6].

MiRTarBase and TarBase are the most popular databases of experimentally validated miRNA: mRNA interactions. These two databases contain comprehensive data of miRNA:mRNA interactions from both dependent and independent sources, and can be downloaded without any restrictions. For these two reasons, both of the databases were selected as the source of experimentally validated miRNA:mRNA interactions. The experimental methods referenced in mirTarbase and TarBase include about 30 different types, such as: reporter assay, western blot, Cross-Linking Immunoprecipitation (CLIP), etc.

Predicted miRNA:mRNA interactions were downloaded from the web servers of miRanda, miRDB, PITA, and TargetScan. The reasons for choosing these four computational predictors are: (1) These predictors are well-designed miRNA target predictors using synthesized techniques. The original version of miRanda predicts miRNA targets based on sequence match, free energy calculated from Vienna RNA package, and evolutionary conservation verified from sequence alignment [3]. Nonetheless, the newest version of miRanda has integrated an mirSVR score into the output prediction [7]. The mirSVR score is the output of “a support vector regression approach to model the degree of microRNA regulation given a set of numerical features representing the microRNA binding site and additional contextual information" [7]. The mirSVR score is actually a measure of the changes of logarithm-based expressional levels of down-regulated mRNAs upon miRNA transfection [7]. The results in miRDB were predicted using mirTarget2, a support vector machine based predictor [8]. The final score of miRDB ranges from 0 to 100 and shows the relative significance of predicted target genes [9]. PITA predicts targets based on sequence match between mRNA and miRNA calculated from RNAduplex [10], and secondary structure of mRNA predicted using RNAFold [10]. The PITA prediction score is an energetic score showing the free-energy change upon the miRNA:mRNA binding [5]. The predictive results of PITA can be further filtered to keep conserved sequences by using phastCons, which is built on hidden Markov model [5,11]. In this study, conserved PITA predictions were used and analyzed. TargetScan was designed to search for the conserved sequence complementarity between the seed region of miRNA and the 3'-UTR of mRNA [6]. The Pct score (Probability of Conserved Targeting) [12] from TargetScan was used to evaluate the significance of the prediction. The default threshold values of positive prediction are $<$-1.0, $>$80, $<$-10, and $>$0.36, for miRanda, miRDB, PITA, and TargetScan, respectively. And vice versa for negative predictions. (2) These predictors were developed or upgraded in recent years, and are well-maintained; (3) These predictors have been broadly used in this field; (4) The web servers of these predictors provide pre-assembled dataset of predicted miRNA:mRNA interactions of mouse genome.

For each pair of miRNA and mRNA molecules, the above-mentioned predictors may make either a valid prediction or an invalid prediction. The difference between valid prediction and invalid prediction is that there is no meaningful output in an invalid prediction, while the valid prediction is always accompanied by a determinative conclusion of the prediction. For this reason, all the predicted miRNA:mRNA interactions download from miRanda, miRDB, PITA, and TargetScan web servers were examined to keep only valid predictions. Since each of these valid predictions could be four-predictor-overlapped or 3-predictor-overlapped or two-predictor-overlapped, this valid prediction was then saved in a corresponding dataset based on which predictors have valid predictions. There are in total $C(4,4) + C(4,3) + C(4,2) = 1 + 4 + 6 = 11$ datasets as shown in Table 1. It should be noted that the number of miRNA:mRNA interaction pairs that can be validly predicted by only one predictor is very limited and therefore these interactions are not included in this study. Afterwards, for each of the afore-mentioned eleven datasets, all the miNRA:mRNA pairs in that dataset were compared with the miRNA:mRNA pairs in the mirTarBase and TarBase databases. If the miRNA:mRNA pair can be found in mirTarBase or Tarbase, the miRNA;mRNA pair is assigned as a true sample. If the pair can't be found in mirTarBase and TarBase, the prediction scores of this miRNA:mRNA pair given by corresponding predictors were compared to their default threshold values. If all the predictions are negative, this miRNA:mRNA pair is assigned as false sample. Otherwise, the miRNA:mRNA pair is excluded from

Table 1
Summary of the eleven datasets.

Dataset ID	Associated predictors	No. of true samples	No. of false samples
D4	miRanda, miRDB, PITA, TargetScan	1739	9870
D3-1	miRanda, miRDB, PITA	1423	11,062
D3-2	miRanda, miRDB, TargetScan	195	2667
D3-3	miRanda, PITA, TargetScan	2997	43,082
D3-4	miRDB, PITA, TargetScan	1990	24,088
D2-1	miRanda, miRDB	187	2722
D2-2	miRanda, PITA	3817	177,467
D2-3	miRanda, TargetScan	423	21,858
D2-4	miRDB, PITA	1713	14,264
D2-5	miRDB, TargetScan	1638	189,900
D2-6	PITA, TargetScan	6211	1340
TOTAL		22,333	498,320

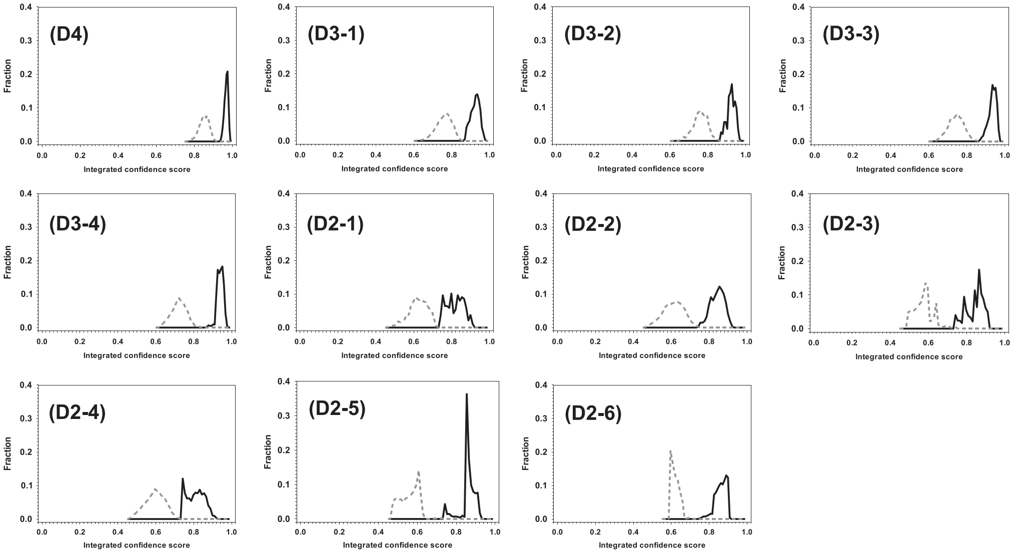

Fig. 1. Distribution of both true and false samples as a function of integrated confidence score in eleven datasets. X-axes show the integrated confidence score. Y-axes present the fractions of true samples (solid lines) and false samples (dashed lines) in each interval of integrated confidence score.

further analysis. The eleven datasets, and corresponding numbers of true and false samples of these eleven datasets are shown in Table 1.

All the miRNA:mRNA pairs in each of the eleven datasets have a specific number of predictive scores given by specific predictors. Therefore, for a miRNA:mRNA pair in a specific dataset, a confidence score was calculated by the following procedure: Assume the i-th sample in this dataset has S predictive scores v_i^s, s=1,…, S. S is the total number of predictors associated with this dataset. By using the predictive scores from the s-th predictor for all the samples in this dataset, the z-scores for all the samples were calculated. Then the p-values were calculated using scipy package of python. Afterwards, the i-th sample in the dataset has S p-values, p_i^s, s=1,…,S. Each of these p-values was then transformed to $d_i^s = 1/(1+e^{-p_i^s})$. Finally, if the sample is a true sample, the integrated confidence score is calculated using $C_i = 1-\prod_1^S(1-d_i^S)$; if the sample is a false sample, the integrated confidence score is determined by $C_i = 1-\prod_1^S d_i^S$. The distributions of integrated confidence scores for both true and false samples in the eleven datasets were presented in Fig. 1. Clearly, different datasets have different ranges of independent variables and dependent variables. For the purpose of showing the difference, we didn't pursue the renormalization of the distribution. A subset of these datasets was refined and applied in one of our recent studies on decision-tree based new algorithm of miRNA target prediction [13].

Acknowledgements

The authors acknowledge with thanks the application of the following online resources: miRanda, miRDB, PITA, TargetScan, mirTarBase, and TarBase.

Transparency document. Supporting information

Transparency data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.07.035.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.07.035.

References

[1] C.H. Chou, N.W. Chang, S. Shrestha, S.D. Hsu, Y.L. Lin, W.H. Lee, C.D. Yang, H.C. Hong, T.Y. Wei, S.J. Tu, et al., miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database, Nucleic acids Res. 44 (D1) (2016) D239–D247.

[2] I.S. Vlachos, M.D. Paraskevopoulou, D. Karagkouni, G. Georgakilas, T. Vergoulis, I. Kanellos, I.L. Anastasopoulos, S. Maniou, K. Karathanou, D. Kalfakakou, et al., DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mrna interactions, Nucleic acids Res. 43 (Database issue) (2015) D153–D159.

[3] A.J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, D.S. Marks, MicroRNA targets in Drosophila, Genome Biol. **5** (1) (2003) R1.

[4] N. Wong, X. Wang, miRDB: an online resource for microRNA target prediction and functional annotations, Nucleic acids Res. 43 (Database issue) (2015) D146–D152.

[5] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, E. Segal, The role of site accessibility in microRNA target recognition, Nat. Genet. 39 (10) (2007) 1278–1284.

[6] B.P. Lewis, C.B. Burge, D.P. Bartel, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets, Cell 120 (1) (2005) 15–20.

[7] D. Betel, A. Koppal, P. Agius, C. Sander, C. Leslie, Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites, Genome Biol. 11 (8) (2010) R90.

[8] X. Wang, I.M. El Naqa, Prediction of both conserved and nonconserved microRNA targets in animals, Bioinformatics 24 (3) (2008) 325–332.

[9] X. Wang, Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies, Bioinformatics 32 (9) (2016) 1316–1322.

[10] R. Lorenz, S.H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, I.L. Hofacker, ViennaRNA Package 2.0, Algorithms Mol. Biol.: AMB 6 (26) (2011).

[11] A. Siepel, D. Haussler, Phylogenetic hidden Markov models, in: R. Nielsen (Ed.), Statistical Methods in Molecular Evolution, Springer-Verlag, New York, 2005, pp. 325–351.

[12] R.C. Friedman, K.K. Farh, C.B. Burge, D.P. Bartel, Most mammalian mRNAs are conserved targets of microRNAs, Genome Res. 19 (1) (2009) 92–105.

[13] B. Zhao, B. Xue, Improving prediction accuracy using decision-tree-based meta-strategy and multi-threshold sequential-voting exemplified by miRNA target prediction, Genomics 109 (3–4) (2017) 227–232.