miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions

Chih-Hung Chou1,2,†, Sirjana Shrestha2,†, Chi-Dung Yang2,3,†, Nai-Wen Chang4, Yu-Ling Lin2,5, Kuang-Wen Liao2,6, Wei-Chi Huang1,2, Ting-Hsuan Sun2, Siang-Jyun Tu1, Wei-Hsiang Lee2,7, Men-Yee Chiew2, Chun-San Tai2, Ting-Yen Wei8, Tzi-Ren Tsai2, Hsin-Tzu Huang2, Chung-Yu Wang6, Hsin-Yi Wu6, Shu-Yi Ho2, Pin-Rong Chen6, Cheng-Hsun Chuang6, Pei-Jung Hsieh2, Yi-Shin Wu2, Wen-Liang Chen2, Meng-Ju Li2,9, Yu-Chun Wu2, Xin-Yi Huang2, Fung Ling Ng2, Waradee Buddhakosai2, Pei-Chun Huang10, Kuan-Chun Lan6, Chia-Yen Huang2, Fung Ling Ng2, Waradee Buddhakosai2, Pei-Chun Huang10, Kuan-Chun Lan6, Chia-Yen Huang2,11, Shun-Long Weng12,13,14, Yeong-Nan Cheng1, Chao Liang1, Wen-Lian Hsu15 and Hsien-Da Huang1,2,16,17,*

1Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, 300, Taiwan, 2Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, 300, Taiwan, 3Institute of Population Health Sciences, National Health Research Institutes, Miaoli, 350, Taiwan, 4Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, 106, Taiwan, 5Center for Bioinformatics Research, National Chiao Tung University, Hsinchu, 300, Taiwan, 6Institute of Molecular Medicine and Bioengineering, National Chiao Tung University, Hsinchu, 300, Taiwan, 7Clinical Research Center, Chung Shan Medical University Hospital, Taichung, 402, Taiwan, 8Interdisciplinary Program of Life Science, National Tsing Hua University, Hsinchu, 300, Taiwan, 9Department of Pediatrics, National Taiwan University Hospital Hsinchu Branch, Hsinchu, 300, Taiwan, 10Delivery Room, Department of Nursing, National Taiwan University Hospital Hsinchu Branch, Hsinchu, 300, Taiwan, 11Gynecologic Cancer Center, Department of Obstetrics and Gynecology, Cathay General Hospital, Taipei, 106, Taiwan, 12Department of Obstetrics and Gynecology, Hsinchu Mackay Memorial Hospital, Hsinchu, 300, Taiwan, 13Mackay Medical, Nursing and Management College, Taipei, 112, Taiwan, 14Department of Medicine, Mackay Medical College, New Taipei City, 252, Taiwan, 15Institute of Information Science, Academia Sinica, Taipei, 115, Taiwan, 16Warshel Institute for Computational Biology, The Chinese University of Hong Kong (Shenzhen), Shenzhen, Guangdong Province, 518172, China and 17School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Shenzhen, Guangdong Province, 518172, China

Received September 15, 2017; Revised October 13, 2017; Editorial Decision October 13, 2017; Accepted October 25, 2017

ABSTRACT

MicroRNAs (miRNAs) are small non-coding RNAs of ~22 nucleotides that are involved in negative regulation of mRNA at the post-transcriptional level. Previously, we developed miRTarBase which provides information about experimentally validated miRNA-target interactions (MTIs). Here, we describe an updated database containing 422,517 curated MTIs from 4076 miRNAs and 23,054 target genes collected from over 8500 articles. The number of MTIs curated by strong evidence has increased ~1.4-fold since the last update in 2016. In this updated version, target sites validated by reporter assay that are available in the literature can be downloaded. The target site sequence can extract new features for analysis via a machine learning approach which can help to evaluate the performance of miRNA-target prediction tools. Furthermore, different ways of browsing enhance user browsing specific MTIs. With these improvements, miRTarBase serves as more comprehensively annotated, experimentally validated miRNA-target interactions databases in the field of miRNA related research. miRTarBase is available at http://miRTarBase.mbc.nctu.edu.tw/.

INTRODUCTION

MicroRNAs (miRNAs) are, small non-coding RNAs with 18–25 nucleotides, which are central regulators at the post-
transcriptional level in both animals and plants. Perfect or near-perfect complementary binding of miRNAs and their target mRNA negatively regulates gene expression by accelerating mRNA degradation or suppressing mRNA translation (1). miRNAs are involved in many biological processes like cell-cycle (2), cell differentiation (3), and apoptosis (4) as well as in different diseases including cancer (5). The study of miRNAs and their target interactions has been gaining interest among researchers due to their causal relationship with disease development. Importantly, dysregulated expression patterns of miRNA may serve as potential biomarkers for disease diagnosis and prognosis (6).

In order to analyze and identify miRNA-target interactions (MTIs), many web-based miRNA-related databases have been established. Of these, miRBase (5) is the primary miRNA sequence repository that facilitates searches for comprehensive miRNA nomenclature, sequence, and annotation data. The DIANA-TarBase (6) is a database that contains manually curated experimentally validated miRNA-gene interactions, with detailed meta-data, experimental methodologies, and conditions. HMDD (7) and miR2Disease (8) contain experimentally validated human miRNA and disease associated MTIs. Although MTIs databases are widely available, MTI-related research has been prolific in recent years. Thus, a centralized information repository of experimentally validated miRNA-target interactions that is easily accessible and updated over the long term must be developed.

Even though computational approaches, that rely on base pairing of miRNAs and their targets, are useful for identifying miRNAs and their target binding sites (9), perfect seed pairing may not always represent the actual interaction between an miRNA and its target (10). Thus, experimental verification must still be carried out to establish miRNAs and their target interactions to elucidate the functions of a given miRNA. MTIs can be experimentally verified in several ways (11). The expression levels of mRNA can be detected using techniques like qRT-PCR, whereas protein products can be measured by ELISA, immunohistochemistry, and Western blot. Reporter assay is the most reliable method to demonstrate direct interaction between an miRNA and its target by measuring reducing activity or the expression of reporter protein (12). Methods like microarray or high-throughput sequencing provide indirect relationships between miRNAs and their targets (13). The newly developed techniques such as CLIP (14), PAR-CLIP (15) and CLASH (16) incorporate high-throughput sequencing that can also be used to identify miRNAs and their targets.

Recently, miRNA-target interactions related research has rapidly increased (Figure S1). To enhance the identification of miRNA targets and their roles in various biological processes, we previously developed miRTarBase (17–19). To date, it has been widely used for improving the accurate rate of target prediction, integrating miRNA related databases or web-based tools, reconstructing miRNA-gene regulatory networks for different diseases, as well as the incorporation into The Cancer Genome Atlas (TCGA) Research Network study. Herein, we provide an up-to-date repository of experimentally validated MTIs with improved annotations and visualization data to facilitate biological investigation, network-based studies and computational machine learning applications.

SYSTEM OVERVIEW AND DATABASE CONTENT

Since the first release of miRTarBase in 2011 until the present, the body of experimentally supported data content has not only dramatically increased but the various functions of the web interface have also been extended. Here, we briefly describe the data content and web-based functions. The database aims to provide the more comprehensive collection of experimentally supported MTIs either in data content or in the web-based function, to accelerate miRNA research. Figure 1 illustrates the current design and system flow of miRTarBase. First, MTI related articles are downloaded from PubMed. Second, enhanced text mining systems are used to extract MTIs. Third, our curator group manually curates the articles to retrieve the important information regarding experimentally verified miRNA-target interactions. We integrated many databases and standalone tools including miRBase (5) and HMDD (7) for miRNA and disease information, and NCBI Entrez gene (20) and RefSeq (21) for target gene information and 3’ untranslated region of target sequences. Likewise, we integrated PubMed, to provide article information, TargetScan (22) and miRanda (23) for target prediction, Gene Expression Omnibus (GEO) (24) and The Cancer Genome Atlas (TCGA) (25,26) for gene and miRNA expression profiling, and KEGG (27) and DAVID (28) for functional annotations of miRNA target genes. Independent studies of MTIs usually describe miRNA using old versions of their name or do not clearly describe the mature form of miRNA. To overcome this problem, our curators checked for the mature form of sequence or confirmed the miRNA version of each study. We also uniformly mapped the miRNA accession to the miRBase v21; unmapped miRNAs were excluded from our database. In addition, gene names were also uniformly mapped by Entrez gene symbol or by Entrez gene alias name (downloaded August 2017). Datasets from the GEO and TCGA with miRNA and mRNA expression profiles were incorporated into miRTarBase. Currently, the database includes 21 human datasets from the NCBI GEO database with at least nine matched mRNA and miRNA samples containing 1596 samples. In addition, a total of 571 samples of 19 cancers from two platforms of TCGA were integrated into the current database. As a result, correlations of expression between miRNA and mRNA in matched samples provide further evidence for direct targets of miRNA.

UPDATED DATABASE CONTENT AND STATISTICS

The improvements and updated database content that are provided by the miRTarBase 2018 are presented in Table 1. The major improvements in this version include (i) a significant increase in the number of curated articles and MTIs, supported by both strong and limited experimental evidences; (ii) the extraction of miRNA-target interactions using an enhanced automatic text-mining system for further manual curation and (iii) the establishment of an enhanced web interface for miRNA-target sites and disease information. The improvements in this version are described in detail as follows.
Figure 1. Overall system flow of miRTarBase. miRTarBase aims to be a more comprehensive database of experimentally verified miRNA-targets interaction. This database was released in 2011. The current database contains more than 420,000 MTIs.

Table 1. Advances, improvements and the number of miRNA-target interactions with different validation methods provided by miRTarBase 7.0

<table>
<thead>
<tr>
<th>Features</th>
<th>miRTarBase 6.0</th>
<th>miRTarBase 7.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Release date</td>
<td>2015/09/15</td>
<td>2017/09/15</td>
</tr>
<tr>
<td>Known miRNA entry</td>
<td>miRBase v20</td>
<td>miRBase v21</td>
</tr>
<tr>
<td>Known Gene entry</td>
<td>Entrez 2015</td>
<td>Entrez 2017</td>
</tr>
<tr>
<td>Species</td>
<td>18</td>
<td>23</td>
</tr>
<tr>
<td>Curated articles</td>
<td>4966</td>
<td>8510</td>
</tr>
<tr>
<td>miRNAs</td>
<td>3786</td>
<td>4076</td>
</tr>
<tr>
<td>Target genes</td>
<td>22 563</td>
<td>23 054</td>
</tr>
<tr>
<td>CLIP-seq datasets</td>
<td>138</td>
<td>231</td>
</tr>
<tr>
<td>Curated miRNA-target interactions</td>
<td>366 181</td>
<td>422 517</td>
</tr>
<tr>
<td>Text-mining technique to prescreen literature</td>
<td>NLP</td>
<td>Enhanced NLP</td>
</tr>
<tr>
<td>Download by validated miRNA-target sites</td>
<td>None</td>
<td>Yes</td>
</tr>
<tr>
<td>Browse by miRNA, gene, and disease</td>
<td>None</td>
<td>Yes</td>
</tr>
<tr>
<td>MTIs Supported by strong experimental evidences</td>
<td>6694</td>
<td>9489</td>
</tr>
<tr>
<td>Number of MTIs validated by ‘Reporter assay’</td>
<td>4580</td>
<td>7258</td>
</tr>
<tr>
<td>Number of MTIs validated by ‘Western blot’</td>
<td>4645</td>
<td>8210</td>
</tr>
<tr>
<td>Number of MTIs validated by ‘Reporter assay and Western blot’</td>
<td>3854</td>
<td>6032</td>
</tr>
<tr>
<td>Number of MTIs validated by ‘Reporter assay or Western blot’</td>
<td>7439</td>
<td>10 581</td>
</tr>
</tbody>
</table>

This version has added a significant number of highly curated articles and expanded its content. The current release (September 2017, version 7) includes a total 422 517 curated MTIs between 4076 miRNAs and 23 054 target genes, which were collected from 8573 articles. The number of MTIs has increased ~1.4-fold curated by strong evidence (Reporter assay or Western blot) since the 2016 miRTarBase update. MTIs which are supported by different experimental approaches are currently available. The update has significantly increased the number of MTIs in comparison to miRTarBase 6.0, as shown in Table 1. The top 10 miRNAs with different target genes on strong evidence are shown in Figure 2A. The top 10 genes, which were regulated by different miRNA on strong evidence, are shown in Figure 2B. The top 10 human disease of different literatures are shown in Figure 2C. The relevant literature was manually surveyed.
to ensure collection of strong evidence for MTIs. As a manually curated, experimentally supported database, the updated version retrieves MTIs that are supported by strong evidence. For example, we have curated roughly 3,000 new MTIs supported by reporter assay and/or Western blot. In addition, many MTIs were supported by more than one independent study. The CLIP-seq datasets include HITS-CLIP and PAR-CLIP, which provide limited experimental evidence and were analyzed as described below.

Collection and analysis of the newly published CLIP-seq data sets

The previous version of database contained a total of 138 CLIP-seq datasets from 21 independent studies. This updated version contains additional 93 CLIP-seq data from four independent studies (29–32). The samples were annotated using the same categories as used previously: source of dataset, types of NGS approach and RNA binding protein (RBP), target species, dataset accession number, tissue or cell line, condition (Supplementary Table S1). CLIP-seq raw data were analyzed using miRTarCLIP (33). In addition, Piranha (34) was used for peak calling and TargetScan (Release 7.1) (22) was used to identify miRNA target sites. The detailed miRTarCLIP process involves trimming adapter reads, removing low-quality reads (phred quality score ≥20 were retained), cytosine (C) to thymine (T) conversion (only in PAR-CLIP data), aligning reads against the reference 3'UTR sequence with TargetScan, clustering target sites and peak calling with Piranha, and finally analyzing miRNA-target sites. As a result, the number of MTIs has been significantly increased; there was an ~1.4-fold increase in MTIs that were validated by NGS, and these are shown in Table 1.

Literature associated with disease

Most of the literature, especially more recent studies, indicates that MTIs are associated with human diseases. Some studies have reported that circulating miRNAs function as biomarkers for disease diagnosis and provide indications of potential target therapeutic agents. In previous version, we integrated data from HMDD v2.0 (7) to represent dynamic miRNA-disease relationships with word clouds. Here, we provide an ‘Evidence’ section containing disease information from the literature on each MTI. The browser page also provides a ‘browse by disease’ option. This extensive disease information enables researchers to search for miRNA-associated diseases more quickly and conveniently. For the disease annotation task, here we collected a list of disease name based on MeSH tree and adopted a partial matching algorithm. After proofreading, all of the extracted articles contained full or part of the disease name.

ENHANCED TEXT MINING SYSTEM

Accumulating knowledge regarding MTIs in the literature can expedite research aimed at understanding the complex physiological processes between miRNAs and their targets. However, the large volume of published work poses a formidable challenge. We devised an automatic MTI detection pipeline comprised of various text processing components. In essence, the MTI system inspects the literature, typically obtained from large databases such as PubMed, and identifies pairs of miRNA and the target gene pairs therein. Our system encompasses two key components: named entity recognition (NER) and relation extraction (RE), as well as other common natural language processing (NLP) procedures. More specifically, the NER module identifies entities mentioned in the text and labels them with appropriate tags. Subsequently, the RE module detects potential relations among those entities using information provided in the previous step. Existing work regarding these tasks can be broadly distinguished into two categories: rule-based and machine-learning-based approaches, and have been widely adopted to conduct MTI on public miRNA databases including DIANA-TarBase v7.0 (6), miRCancer (35) and miRSel (36).

However, there are some limitations and weaknesses inherent to both approaches. In rule-based methods, explicit rules must be developed by domain experts but it is hard to cover all the possible variations with manually defined rules. Thus, inflexible rules make it difficult for approaches to deal with unpredictable exceptions, which may result in some types of word mutations, such as insertion, deletion, or substitution (IDS), in the target named entities. In contrast, machine learning (ML) models can automatically learn implicit patterns for generating principles. Nevertheless, the...
ML-based model is not as intuitive as rule-based methods and the results cannot be easily interpreted by humans. Consequently, we utilized the statistical-principle-based approach (SPBA) (37) to implement our MTI detection system. Figure 3 illustrates the details of the processing in our MTI detection system.

On the whole, the developed approach, SPBA, can automatically elicit labeled sequences and incorporate them into more representative principles by observing dominated principles. Afterwards, a partial matching algorithm is employed in harnessing the advantages of both rule- and machine learning-based approaches while surpassing their limitations. The method is proved to outperform in several distinct tasks such as topic detection and sentimental analysis (37–40).

In this updated version, a set of principles generated in the previous version was manually curated by domain experts to enhance the recognition of miRNA-target interactions (MTIs). Afterwards, the proposed SPBA-based approach can successfully extract the relation between miRNA and their targets from articles by the partial matching algorithm. All of the selected articles in this update are validated by several curators. Through the screening mechanism based on SPBA, the enhanced text mining pipeline singled out 3371 papers, which can be referred to as a summary of MTIs. After manual proofreading, 2617 papers were confirmed to contain MTIs, indicating that precision of 78% was achieved by our system. The enhanced SPBA-based MTI detector improved on the precision of the previous version by 7%.

ENHANCED WEB INTERFACE

The previous version provided a user friendly, graphical visualization interface for biologists to investigate MTIs in detail. We provide all of the possible relevant information, including miRNA secondary structure, miRNA-related diseases, putative target sites, target sites provided by the authors, evidence of MTIs from original descriptions in each article, expression profiling of miRNAs and their target genes using GEO and TCGA datasets, CLIP-seq evidence for MTIs, target gene set enrichment analysis, and the MTI networks. However, improvements were required to identify easier ways to find the relevant MTIs. Here, we provide three different ways to browse for specific MTIs: 'browse by miRNA', 'browse by target genes', and 'browse by disease'. Upon browsing, these three categories present the top 50 results. Based on text-mining, a new column for disease description has been added to each MTI in the 'Evidence' page, which provides information regarding the experimentally supported literature. The most important new element is that the browser provides the target site in web interface and also downloads pages for further study. In the era of artificial intelligence, especially with regards to machine learning and deep learning applications, we believe that the datasets retrieved from experimentally validated miRNA-target sites improve the accuracy of current target predic-
tion tools and reduce the chances of false-positive predictions.

DATA AVAILABILITY

The Download page of miRTarBase provides the more comprehensive data for download that include: (1) All published miRNA target interaction data. (2) MTI PubMed Abstract Manually Curation Corpus (3) Catalog by Species of MTIs. (4) Catalog by Experimental Evidences of MTIs. We also provide all the previous version release in the ‘Previous Release’ page.

DISCUSSION

We aimed to provide more convenient and accurate information regarding MTIs, speeding up miRNA research in biomedical science, and provide assistance to researchers in the field. Since 2011, many researchers have used miRTarBase to conduct fruitful research, including improving the accuracy rate of target prediction, integrating miRNA related databases and web-based tools, and reconstructing miRNA-gene regulatory networks in different diseases. In addition, the database was incorporated into The Cancer Genome Atlas Research Network study. The main applications of the database are described below as well as in Supplementary DATA and Table S2 in detail.

miRTarBase has recently been used to conduct integrative analysis in many high quality miRNA related studies, such as clarifying the super-enhancer-mediated RNA processing mechanism (41), informing the miRNA dysregulation in autism spectrum disorder (ASD) and neurological disease (42), supporting the design of tumor type-specific treatments (43), and expanding miRNA therapeutics, which is crucial to disease treatment and cure (44). Moreover, our database also functions significantly in research on different types of cancer, including liver cancer (45), ovarian cancer (46) and gastric cancer (47). Remarkably, in The Cancer Genome Atlas (TCGA) project, the research team integrated high-throughput experimental data and miRTarBase’s MTI information to accomplish important research work and reveal a complex molecular landscape of cervical cancers and Pancreatic Ductal Adenocarcinoma (PDAC) to provide a roadmap for precision medicine (48,49).

SUMMARY AND PERSPECTIVES

Since the miRTarBase debuted in 2011, MTIs and related databases have been continually updated. Moreover, the collection of its datasets has been expanded. We have curated over 8,500 experimentally supported articles regarding miRNA-target interactions. As a result, ~10 000 MTIs that have been validated by strong evidence have been included. Our improved text mining pipeline and a set of patents that have been validated by strong evidence have been incorporated. Our improved text mining pipeline and a set of patents that have been validated by strong evidence have been incorporated.

We believe that the target site sequence could extract new features for analysis via a machine learning approach that can help to evaluate the performance of miRNA-target prediction tools. We will extend and expand the miRTarBase continuously to the more comprehensive database by improving NLP technology, recruiting more curators, and enhancing the web function and annotation. To conclude, we provide an extensive repository of experimental information that clearly makes a continuous and worthwhile contribution to miRNA research related to cancer mechanisms, disease diagnosis and treatment.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING


REFERENCES


