miRDB: an online resource for microRNA target prediction and functional annotations

Nathan Wong1,2 and Xiaowei Wang2,1,*

1Department of Biomedical Engineering, Washington University, St. Louis, MO 63130, USA and 2Department of Radiation Oncology, Washington University School of Medicine, St. Louis, MO 63108, USA

Received August 10, 2014; Revised September 29, 2014; Accepted October 23, 2014

ABSTRACT

MicroRNAs (miRNAs) are small non-coding RNAs that are extensively involved in many physiological and disease processes. One major challenge in miRNA studies is the identification of genes regulated by miRNAs. To this end, we have developed an online resource, miRDB (http://mirdb.org), for miRNA target prediction and functional annotations. Here, we describe recently updated features of miRDB, including 2.1 million predicted gene targets regulated by 6709 miRNAs. In addition to presenting precompiled prediction data, a new feature is the web server interface that allows submission of user-provided sequences for miRNA target prediction. In this way, users have the flexibility to study any custom miRNAs or target genes of interest. Another major update of miRDB is related to functional miRNA annotations. Although thousands of miRNAs have been identified, many of the reported miRNAs are not likely to play active functional roles or may even have been falsely identified as miRNAs from high-throughput studies. To address this issue, we have performed combined computational analyses and literature mining, and identified 568 and 452 functional miRNAs in humans and mice, respectively. These miRNAs, as well as associated functional annotations, are presented in the FuncMir Collection in miRDB.

INTRODUCTION

MicroRNAs (miRNAs) are a family of small non-coding RNAs that play important regulatory roles in many physiological and disease processes (1,2). About 2000 human miRNAs have been discovered to date, and collectively these miRNAs regulate the expression of thousands of protein-coding gene targets at both post-transcriptional and translational levels (3–5). miRNAs exert their functions via target downregulation. Thus, identification of gene targets is critical for functional characterization of miRNAs. Currently, experimental identification of miRNA targets is a time-consuming process and as a result most researchers rely on computational tools to initially identify a set of candidate gene targets for further experimental characterization. To facilitate the process of selecting functional targets at the genome level, we have previously developed an online resource, miRDB for miRNA target prediction and functional annotations (6). Here, we present major updates to miRDB, including updated target prediction data, a new web server interface for custom target prediction, as well as the inclusion of a set of functional miRNAs annotated by integrating computational analyses with literature mining. The web interface of miRDB has also been updated to accommodate these new database features at http://mirdb.org/.

METHODS AND RESULTS

Updates on miRNA target prediction data

Since the first version of miRDB was established in 2008, thousands of novel miRNAs have been discovered. In addition, annotations of gene targets, especially concerning the identification of 3′-UTR sequences, have been significantly expanded. Thus, we have performed a major update on target prediction data by employing the up-to-date miRNA and target gene annotations. All mRNA sequences and annotations were downloaded from miRBase (version 21) in June 2014 (7). We have adopted the NCBI RefSeq database for identification of 3′-UTR sequences. In brief, RefSeq sequences were downloaded from NCBI’s ftp site (8) and further parsed with the BioPerl program to obtain the 3′-UTR sequences of the transcripts. Target prediction was then performed with the MirTarget algorithm, which was developed by analyzing high-throughput expression profiling data in a support vector machine framework (9). Unlike most other prediction algorithms, MirTarget predicts both conserved and nonconserved gene targets by treating target site conservation as an important but non-required sequence feature. The robust performance of MirTarget has been extensively demonstrated. For example, a recent independent analysis shows that MirTarget has superior performance over other public algorithms for identifying miRNA-
downregulated gene targets (10). In this miRDB update, we have also updated the MirTarget algorithm by including additional model training data, which were generated from miRNA-target pairs experimentally identified by RNA-seq (11). Details of the algorithmic improvement will be described elsewhere.

With updated genomic data and the MirTarget algorithm, we have performed genome-wide miRNA target prediction for all known transcripts (including all isoforms) from five species—human, mouse, rat, dog and chicken. In total, 2.1 million gene targets were predicted to be regulated by 6709 miRNAs in these five species. All the targets have a prediction score in the range of 50–100 as assigned by MirTarget, with a higher score representing more statistical confidence in the prediction result. Detailed statistics of the target prediction are presented in the miRDB website. All the target prediction data as well as the associated genomic annotations were imported into a backend MySQL database for web presentation. The users can search for precompiled results via miRDB web interface, using either miRNA or gene target search terms. Notably, the users precompiled results via miRDB web interface, using either

One major update of miRDB is the implementation of the miRDB web interface for custom miRNA target prediction. With this new function, users are now able to provide their own sequences (either a miRNA or target sequence) for target prediction. Specifically, the user can choose between providing a miRNA sequence (17–30 nt) or a target gene sequence (100–30 000 nt) for prediction of miRNA targets in one of the five species—human, mouse, rat, dog or chicken. On submission, the server calls a backend Perl script which implements the MirTarget prediction algorithm (Figure 2A). The script takes as its inputs the user-provided sequence, as well as the selected species in which the prediction is to be conducted. For miRNA submission, the server imports all the species-specific miRNAs, as collated from miRBase v21.

For either custom miRNAs or custom targets, the program first evaluates each potential miRNA-target pair by initially screening for 7-mer seed complementarity in the target site. Based on the relative strength of each sequence-based target recognition feature, the MirTarget algorithm is then used to predict miRNA targets. MirTarget was originally developed for genome-wide target prediction, and it has been customized here to become an integral component of the web server pipeline. When the prediction process is completed, the user progresses to the results page. The prediction data are relayed to a Perl-CGI script for web presentation. In addition, the script also retrieves relevant gene annotations for each target from a backend MySQL database. The web page then displays the results in a manner similar to those produced by a standard miRDB query. Briefly, the scores are sorted in descending order and show the user the ranking of the miRNA-target pair, the score, the names of the miRNA and the target, and in the case of custom miRNA submission, the annotations of the gene targets. Additionally, users have the option of viewing more details of the target prediction, which includes both the miRNA and the target sequences and the locations of seed matches (Figure 2B). Concerning the speed of the prediction process, it takes about 5–10 s per query for genome-wide target prediction for a user-provided miRNA sequence; it takes about 1–2 s per query for a typical user-provided target sequence (~2 kb long).

**FuncMir collection for annotating functional miRNAs**

The discovery of a large number of miRNAs presents grand new opportunities to uncover the complexity of gene expression regulatory networks (12). On the other hand, many of the newly discovered miRNAs are not likely to be functionally relevant, or may even have been false discoveries resulting from low-quality data. For example, an RNA-seq experiment, if not properly performed or analyzed, could potentially produce hundreds of falsely discovered ‘novel’ miRNAs. Unfortunately, many such ‘novel’ miRNAs might have already been included in miRBase, which only implements limited quality control filters and in principle allows the submission of any miRNAs that have been reported in peer-reviewed studies (7). As a result, a large number of miRNAs included in miRBase are not likely to play any functional roles in gene expression regulation. Our inability to distinguish functional miRNAs from non-functional ones could severely compromise both computational and experimental studies on miRNA functions, leading to unnecessary wastes of resources or even obscure real important findings. For example, a functional library screen will be negatively impacted by the inclusion of many falsely discovered non-functional miRNAs. On the computational side, the inclusion of many non-functional miRNAs can severely reduce the significance of real biological findings, such as those on functional miRNAs that regulate competing endogenous RNAs (ceRNAs) (13).
Figure 1. miRDB web interface for miRNA target prediction. (A) Search for miRNA targets by a single miRNA or gene target. (B) Search for miRNA targets by multiple miRNAs or gene targets. (C) A screenshot of target prediction data retrieved from miRDB.
Figure 2. Custom target prediction with user-provided sequences. (A) Workflow of the target prediction pipeline for the miRDB web server. (B) A screenshot of the custom web search interface and a representative result page.
Figure 3. Identification of functional miRNAs by combinatorial analysis of four selection criteria. (A) Composite scores were calculated by summarizing individual criterion scores from four categories. All functional miRNAs have a composite score $\geq 3$. (B) Distribution of functional human miRNAs identified by various numbers of scoring criteria. (C) A screenshot of a FuncMir miRNA homepage. For each functional miRNA, a web page is established by integrating data from various sources, including miRBase, PubMed and NCBI Gene databases. The page for hsa-mir-19b-1 is presented here as an example.
A complete list of both composite scores and individual criterion scores for all the miRNAs is presented in Supplementary File 1. For each functional precursor miRNA, the dominantly expressed mature miRNA was selected. In some cases, a second mature miRNA was selected from the same precursor if its expression level was no less than 10% of the dominant mature miRNA and its read count was no less than 5 reads per million reads as determined by RNA-seq experiments. In this way, 654 and 442 functional mature miRNAs were identified in humans and mice, respectively. All the identified functional miRNAs were included in the FuncMir Collection of miRDB, with one entry homepage for each miRNA. A screenshot of a representative miRNA entry is presented in Figure 3C. The CLASH RNA-seq data (11) were analyzed to evaluate the expression profiles of FuncMir miRNAs. Among all 2588 human mature miRNAs included in miRBase, 322 were detected at least five times per million sequence reads in HEK293 cells. Most of these expressed miRNAs (79.2%) have been included in the FuncMir Collection. In summary, the FuncMir Collection helps to facilitate researchers to focus on functionally relevant miRNAs, especially in experimental screens for miRNA functions or systematic computational analysis at the genome level.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.
ACKNOWLEDGEMENT
We thank Matt Narens for technical assistance.

FUNDING
National Institutes of Health (NIH) [R01GM089784 to X.W.]. Funding for open access charge: NIH [R01GM089784 to X.W.].

Conflict of interest statement. None declared.

REFERENCES