Journal of Biomedical Informatics 44 (2011) 839–847

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

miRWalk – Database: Prediction of possible miRNA binding sites by "walking" the genes of three genomes

Harsh Dweep [1], Carsten Sticht [*,1], Priyanka Pandey, Norbert Gretz

Medical Research Center, Medical Faculty Mannheim, University of Heidelberg, D-68167 Mannheim, Germany

ARTICLE INFO

Article history:
Received 3 December 2010
Available online 14 May 2011

Keywords:
Database
miRNA
Promoter
5′-UTR
CDS
3′-UTR
Target prediction
Validation
Text-mining

ABSTRACT

MicroRNAs are small, non-coding RNA molecules that can complementarily bind to the mRNA 3′-UTR region to regulate the gene expression by transcriptional repression or induction of mRNA degradation. Increasing evidence suggests a new mechanism by which miRNAs may regulate target gene expression by binding in promoter and amino acid coding regions. Most of the existing databases on miRNAs are restricted to mRNA 3′-UTR region. To address this issue, we present miRWalk, a comprehensive database on miRNAs, which hosts predicted as well as validated miRNA binding sites, information on all known genes of human, mouse and rat.

All mRNAs, mitochondrial genes and 10 kb upstream flanking regions of all known genes of human, mouse and rat were analyzed by using a newly developed algorithm named 'miRWalk' as well as with eight already established programs for putative miRNA binding sites. An automated and extensive text-mining search was performed on PubMed database to extract validated information on miRNAs. Combined information was put into a MySQL database.

miRWalk presents predicted and validated information on miRNA-target interaction. Such a resource enables researchers to validate new targets of miRNA not only on 3′-UTR, but also on the other regions of all known genes. The 'Validated Target module' is updated every month and the 'Predicted Target module' is updated every 6 months. miRWalk is freely available at http://mirwalk.uni-hd.de/.

1. Introduction

MicroRNAs (miRNAs) are small, non-coding RNA molecules of 21–25 nucleotides in length that regulate the gene expression by base–pairing with the transcripts of their targets i.e. protein-coding genes, leading to downregulation or repression of the target genes [1]. However, target gene activation has also been described [2]. miRNAs are transcribed from long primary transcript (pri-miRNAs) in the nucleus and processed into characteristic stem-loop precursor miRNAs (pre-miRNAs) by the enzyme Drosha. Then pre-miRNAs are transported into cytoplasm, where they are transformed into small, single-stranded miRNAs with the help of Dicer [3]. One strand of the mature miRNA enters the RNA-induced silencing complex (RISC) and binds to the 3′-untranslated region (3′-UTR) of the target mRNA through imperfect base-pairing. Previously it has been shown that 5′ end of miRNA could be determinant in target repression [4]. The 5′ end sequence of miRNA is called "seed" and has a length of 6–8 nucleotides which is energetically favorable for the miRNA target interaction [5]. Mutations in the seed region of a miRNA sequence leads to an inactive interaction [6]. The binding reduces the expression level of target protein by a number of mechanisms including inhibition of translational initiation [7], inhibition of elongation, and induction of deadenylation which decreases mRNA stability and increases the rate of mRNA degradation [8]. The miRNA gene family is one of the largest in higher eukaryotes: more than 700 miRNAs have been identified in the human genome [9], each of them having the potential to bind to hundreds of transcripts. miRNAs are involved in diverse regulatory pathways [10,11], as well as in disease development, progression, prognosis, diagnosis and evaluation of treatment response [12,13].

Computational prediction of miRNA targets is much more challenging in animals than in plants, because animal miRNAs often perform imperfect base-pairing with their target sites [14], unlike plant miRNAs which almost always bind their targets with near perfect complementarity [15]. In the past years, a large number of target prediction programs and databases on experimentally validated information have been developed for animal miRNAs [5,16–26].

For more than a decade, attempts to study the interaction of miRNAs with their targets were focused to the 3′-UTR region of mRNAs. But recent studies on miRNA-target interaction revealed

* Corresponding author. Fax: +49 621 383 2108.
E-mail addresses: harsh.dweep@medma.uni-heidelberg.de (H. Dweep), carsten.sticht@medma.uni-heidelberg.de (C. Sticht), priyankashahi2001@googlemail.com (P. Pandey), norbert.gretz@medma.uni-heidelberg.de (N. Gretz).
[1] Both authors contributed equally to this work.

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.05.002

a new mode of action of miRNAs by which they may regulate the gene expression by targeting promoter as well as amino acid coding (CDS) regions. Tay et al. demonstrated the existence of many naturally occurring miRNA targets sites of miR-314, miR-296 and miR-470 in the CDS of the genes Nanog, Oct4 and Sox2 [27]. Guang et al. have shown that argonaute proteins can transport classes of small regulatory RNA to distinct cellular compartments to regulate gene expression [28]. In another study, Place et al. have demonstrated that miR-373 targets the promoter sequences of E-cadherin and CSDC2 genes to induce gene expression [29]. On the other hand, a few experiments have indicated possible target sites in the 5′-UTR for e.g. [30]. Thus, it is of paramount importance to design a new approach which can identify putative miRNA binding sites not only in the 3′-UTR region, but also in other regions (promoter, 5′-UTR, and CDS) of a gene.

Here we present miRWalk (http://mirwalk.uni-hd.de/), a comprehensive database that provides predicted as well as validated miRNA binding site information on miRNAs for human, mouse and rat.

2. Materials and methods

2.1. Genome data acquisition

All sequences [mRNAs, 10 kb upstream flanking regions (assumed promoter regions) and mitochondrial genes] and other necessary information (EntrezID, mRNA and CDS length, gene location and definition) of all known genes of human, mouse and rat were downloaded by submitting identifiers i.e. RefSeqID, Ensembl gene IDs and official gene symbols to Entrez at GenBank [31] and Ensembl [32] databases. miRNA sequences and other information (e.g. Sanger name, MIID, genomic location of miRNA, stem loop sequence and other accession numbers like stem loop, and mature sequence) were collected from miRBase [33]. For Comparative analysis, DIANA-microT (version 3.0), miRanda (August 2010), miRDB (April 2009), PicTar (March 2007), PITA (August 2008), RNA22 (May 2008) and TargetScan/TargetScanS (version 5.1) predictions files were downloaded from their databases, while in case of RNAhybrid, the prediction data was generated by running the executable (version 2.1) of RNAhybrid on bwGRID server.

2.2. Compilation and curation of keyword dictionary

The gene and miRNA dictionaries for human, mouse and rat were compiled from several databases: HUGO Gene Nomenclature Committee (HGNC), Mouse Genome Database (MGD), Rat Genome Database (RGD), gene-centered information at NCBI (Entrez Gene), Targetscan and miRBase. The names, aliases, symbols, official names, synonyms and database identifiers were merged into synonym dictionaries. In the curation step, inappropriate synonymous or expressions that would lead to ambiguous or wrong identifications were detected and removed as described [34]. The information on diseases, organs, and OMIM disorders was extracted from MeSH (Medical Subject Heading) and OMIM [35]. Whereas the keywords on proteins known to be involved in miRNA processing and cell lines were collected from PubMed database by reading the publications. Collated information was then organized in several lists. Then, these keywords were compiled, classified and stored in different dictionaries according to disease, organ, cell line,

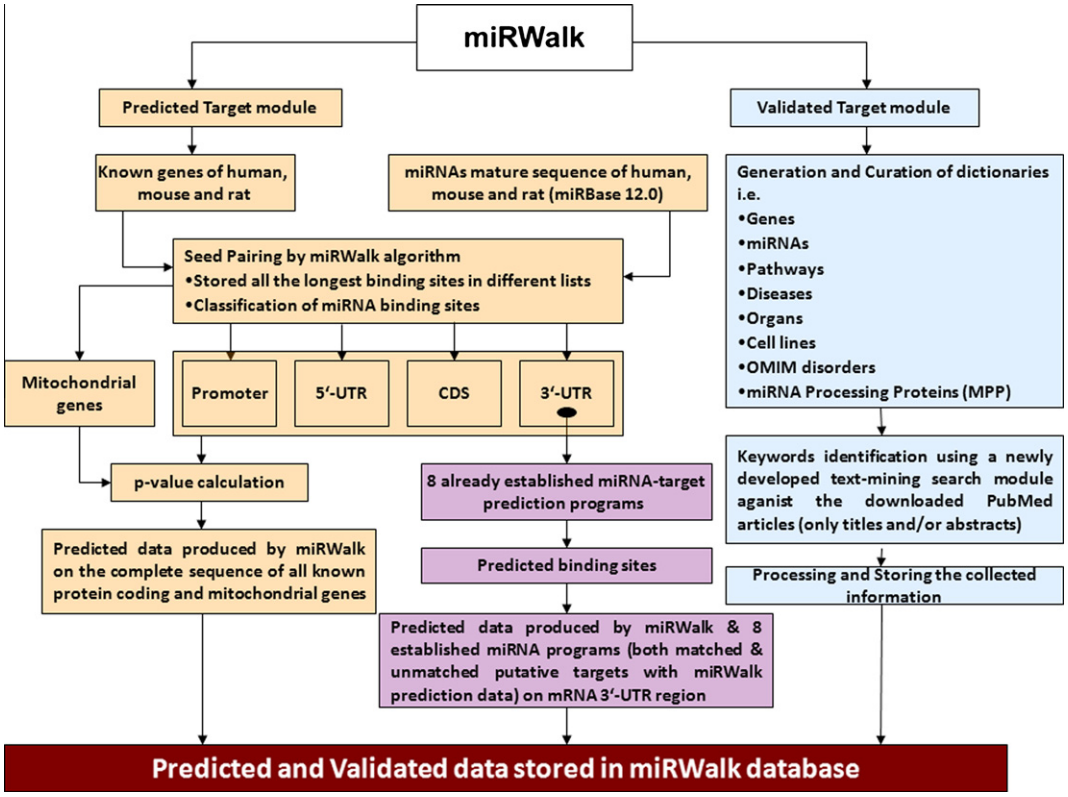

Fig. 1. Workflow of miRWalk algorithm and automated text-mining module. miRWalk consists of two modules, i.e. Predicted Target and Validated Target. 'Predicted Target', miRWalk searches for the longest complementary matches between miRNAs and all downloaded sequences. Afterwards, it classifies all identified hits in protein coding (four regions, i.e. Promoter, 5′-UTR, CDS, and 3′-UTR) and mitochondrial genes. Then the probability distribution of random matches of a subsequence (longest miRNA binding sites) in the analyzed sequence is calculated by using a Poisson distribution. Afterwards, miRWalk compares the identified miRNA binding sites with the results obtained from eight established miRNA-target prediction programs. 'Validated Target', performs an automated text-mining search in the titles/abstracts of the PubMed articles by using curated dictionaries. Finally predicted and validated information is stored in a relational database called 'miRWalk'.

OMIM (Online Mendelian Inheritance in Man) and miRNA processing proteins. Simultaneously, the information on pathways such as gene sets, names and pathway identifiers were retrieved from KEGG [36] and Biocarta (www.biocarta.com) databases. Thereafter, all the available abstracts that contained miRNA keywords in their title and/or abstracts were downloaded from PubMed database.

2.3. miRWalk algorithm

The miRWalk algorithm is based on a computational approach which is written in Perl programming language to identify multiple consecutive Watson–Crick complementary subsequences between miRNA and gene sequences as described in Fig. 1 (Predicted Target module). miRWalk algorithm searches for seeds based on Watson–Crick complementarity, walking on the complete sequence of a gene starting with a heptamer (seven nucleotides) seed from positionS 1 and 2 of miRNA sequences. As soon as it identifies a heptamer perfect base-pairing, it immediately extends the length of the miRNA seed until a mismatch arises. It then returns all possible hits with 7 or longer matches. These binding sites are then separated on the basis of their identified locations (start, and end positions and regions) in the analyzed sequences. Then it assigns the prediction results in five parts, according to promoter region, 5′-UTR, coding sequence (CDS), and 3′-UTR and mitochondrial genes.

In addition, the probability distribution of random matches of a subsequence (5′ end miRNA sequence) in the analyzed sequence is calculated by using a Poisson distribution [20]. It can be expected that the longer perfect complementation of a seed is associated with a lower probability, thus the higher are the chances of an effective miRNA-target interaction.

2.4. Incorporating other prediction databases in miRWalk

Comparative studies conducted with the earlier miRNA target prediction programs suggested that no program was consistently superior to all others [37,38]. Indeed, it has become a common practice for researchers to look at predictions produced by several miRNA-target prediction programs and focus on their intersection [39,40]. Thus, miRWalk compares the identified miRNA binding sites with the results obtained from eight established miRNA-target prediction programs. These eight established programs are chosen on the basis of their popularity. Finally miRWalk incorporates all the predicted miRNA binding sites of the miRWalk algorithm and of the eight programs as predicted information on miRNAs into its database. All the putative targets (both matched as well as unmatched with miRWalk prediction data) of other programs are stored in miRWalk database.

2.5. Automated text-mining search module

A new module was written in Perl programming language to accomplish the automated text-mining task in the downloaded titles and abstracts from PubMed database against curated dictionaries. We chose abstracts because they are not only more readily available than full text, but are also condensed descriptions focusing on what is central to the study, combining the back-

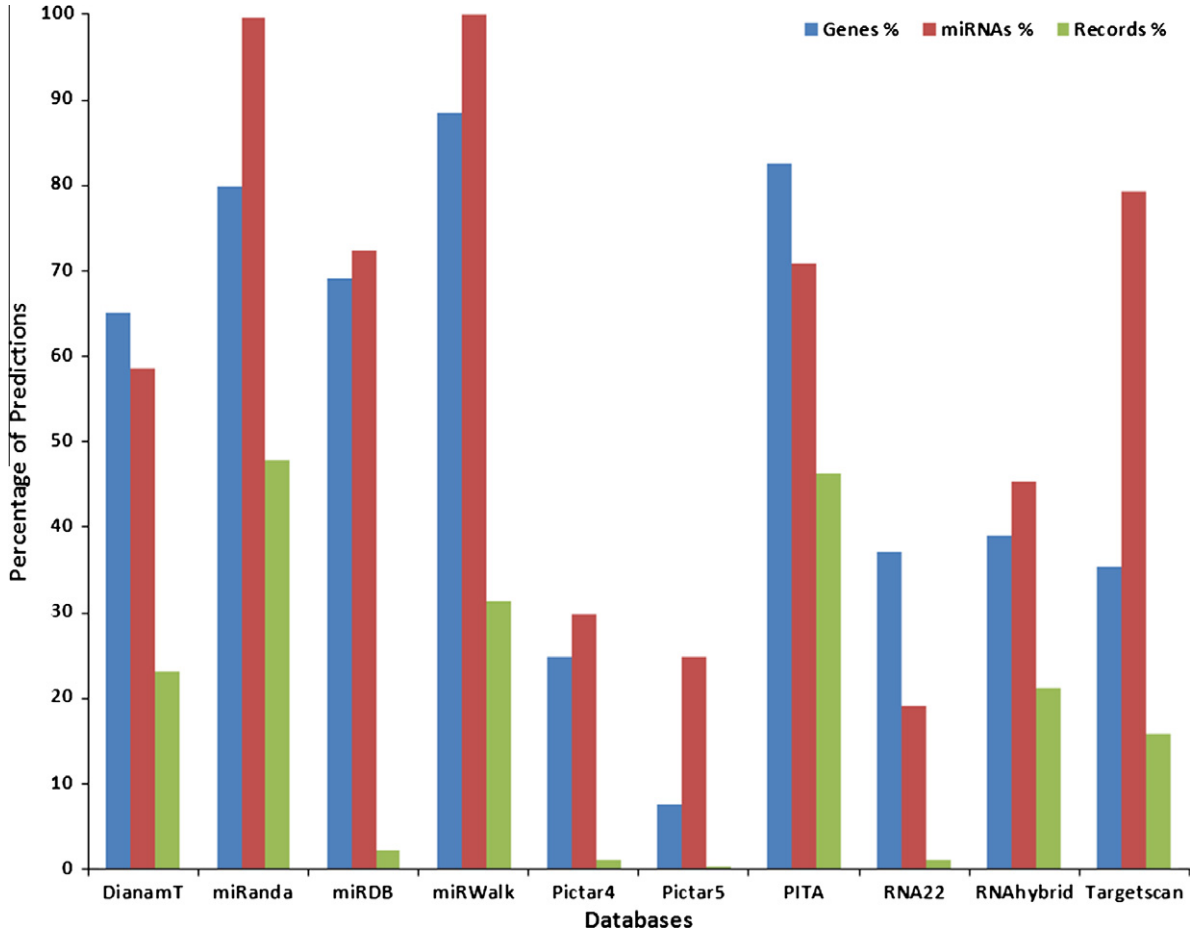

Fig. 2. An overview of predictions within mRNA 3′-UTR region by miRWalk and eight other databases. In promoter region, more than 97% of the genes were identified as the putative targets of one or more miRNAs and covered 72% interactions. Whereas, the coding, 3′ UTR and 5′ UTR regions covered 15%, 10% and 1.6% interactions on 97.8%, 88.5% and 80% of genes, respectively. See Supplementary Table 1 for more details on interaction data on all genes of human, mouse and rat, produced by miRWalk algorithm.

ground, results and conclusions. Shah et al. [41], demonstrated that the abstract section of articles contains the best proportion of keywords, while the other sections are a better source of biologically relevant data. In details, first the automated text-mining search module downloads all the available abstracts in XML (Extensible Markup Language) format that contained keywords such as 'microRNA', 'miRNA', 'micro-RNA', 'micro_RNA', 'micro RNA', 'miR' in their title and/or abstracts. Thereafter, the keywords from curated dictionaries are detected in the texts of downloaded titles/abstracts (Name Entity Recognition, NER) by string matching script with the help of Perl regular expressions. The identification of named entities allows identifying miRNA terms linked with genes, diseases, organs, cell lines, pathways, OMIM disorders and proteins known to be involved in miRNA processing in an abstract. Then these relationships are processed and stored along with their PubMed identifiers as validated information under Validated Target module of miRWalk database as shown in Fig. 1 (Validated Target module).

2.6. Data processing

After sequence acquisition, miRWalk algorithm was executed to produce putative miRNA binding sites on all downloaded sequences. The bwGRID Cluster (high performance cluster facility) has been used to run miRWalk algorithm in a batch mode by adopting four nodes with 32 processors for the faster identification of miRNA binding sites and computing of probability. After the identification of the miRNA binding sites with miRWalk, the predicted miRNA-target interactions data on 3′-UTR of all known genes of human, mouse and rat were obtained from eight established miRNA-target prediction programs for the comparison of the results with different algorithms. The predicted miRNA-target interaction data on more than 2000 miRNAs generated by both miRWalk and eight established miRNA-target prediction programs was documented in the Predicted Target module.

For validated information, the dictionaries for human, mouse and rat were created from several databases as described in compilation and curation section of this paper. In a keyword identification step, a newly developed automated text-mining search module was used to detect gene and miRNA names in the downloaded PubMed abstracts against curated dictionaries by using NER. Then the collated information was processed and stored under Validated Target module of miRWalk database.

3. Results

3.1. miRWalk database and web interface

miRWalk is implemented as a relational database on a MySQL database management system. The web interface of miRWalk database is divided into two modules: (i) Predicted Target module and (ii) Validated Target module.

The Predicted Target module is classified into six parts: Target Gene, miRNA, Pathway, Chromosome, OMIM and Mitochondrial Target. There are two options implemented for user input, i.e. entering the gene symbols or EntrezIDs either directly into a textbox area or by uploading a file under Target Gene and MicroRNA Target search pages. Whereas other predicted search pages can

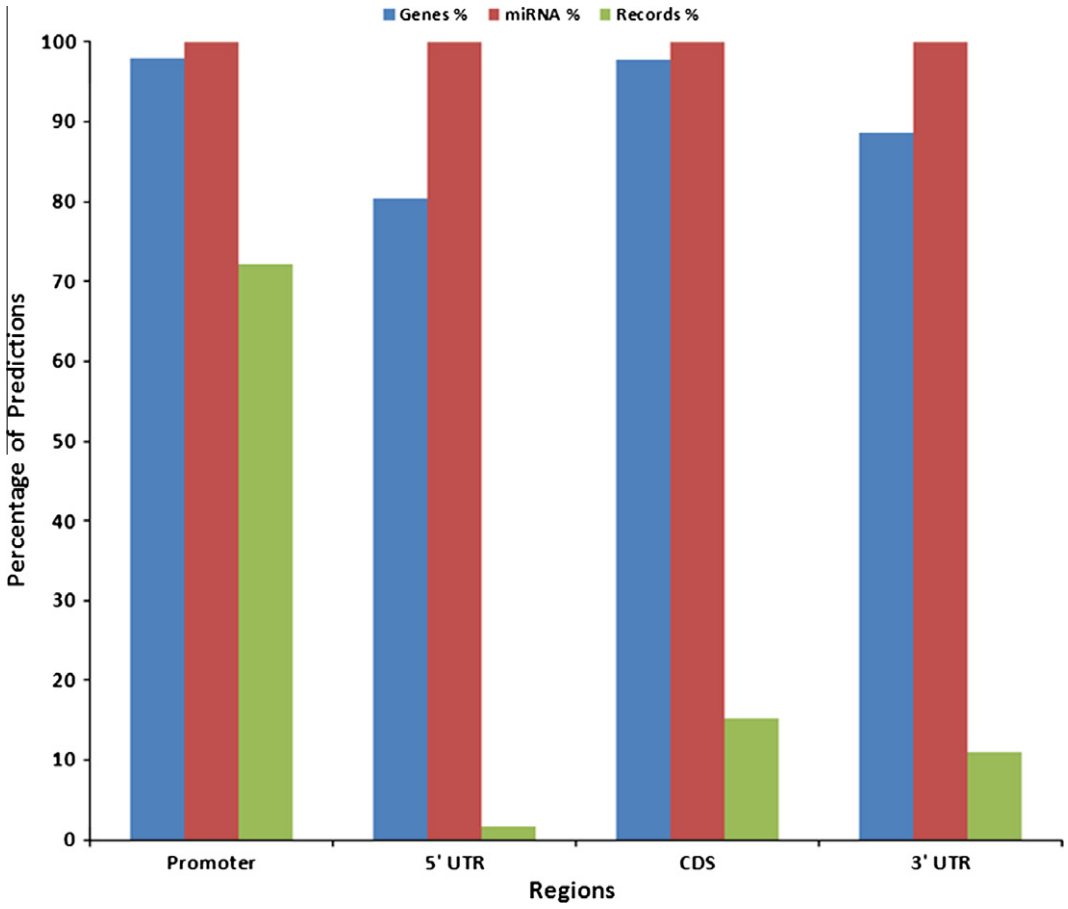

Fig. 3. Overview of the predictions produced by miRWalk algorithm. More information on the estimation of predictions in different databases is given in Supplementary Table 2.

be easily queried by a simple selection of drop down lists for e.g. Pathway Target search page can be easily inquired by choosing a pathway name from the given drop down list, to determine how many genes of a selected pathway are the targets of similar and/or different miRNAs. The result tables of all searches display only 20 records and appropriate links are given, which redirect the user to NCBI, Ensembl, UCSC, miRBase, KEGG, Biocarta and OMIM databases for more annotation and further information on the results. Moreover, one can view and/or download the complete result tables by clicking on the providing links such as 'View Complete Table', 'Paging View' and 'Download Table'.

The Validated Target module has different search pages. These search pages are organized similar to Predicted Target module and are called; Target Gene, miRNA, Pathway, Disease, Organ, Cell line, miRNA literature, OMIM disorder and miRNA Processing Proteins. The results of this module are hyperlinked to PubMed database. More information on web interface can be found in Supplementary material.

3.2. Overview of predictions produced by miRWalk and eight established databases

Since miRNA-target prediction programs are designed with different combinations of features to perform the same task, it is essential to analyse the distribution of these predictions across miRWalk and eight established miRNA-target prediction programs. We count over seven million predictions in both miRanda (36,507 genes and 2050 miRNAs) and PITA (37,714 genes and 1458 miRNAs), whereas over 4 million unique predictions (only single binding site of a miRNA per 3′-UTR region) are counted on 3′-UTR by miRWalk algorithm which cover over 40,491 genes (from all three species) and 2057 miRNAs (Fig. 2). An overview of miRWalk predictions on different regions is presented in Fig. 3. In a comparative study, the predictions are decreasing to a surprisingly small 79 records (73 genes and 34 miRNAs) identified in 10 datasets of nine databases considered (Fig. 4). Fig. 4 depicts overlaps of predictions among nine different databases (including miRWalk predictions on 3′-UTR region) and indicates that although it shows low total overlap among all databases, in spite of that there is considerable similarity between at least five prediction databases.

3.3. Evaluation of miRWalk algorithm

We compared the performance of miRWalk with respect to eight already established programs. For evaluation we selected a set of genes (positive and negative sets) from TarBase (version 5.0) [42], miRecords (version 3) [21] and miRTarBase (release 2.1) [43] databases on which miRNA binding sites are already verified and published in the PubMed database. We analyzed 1870 positive miRNA-target and 61 negative miRNA-target pairs for the performance of different prediction programs. The positive and negative datasets were given as input to each of these programs by using 'Target Gene' which is implemented under Predicted Target module of miRWalk database and the output was analyzed to calculate the Accuracy (Acc), Recall and Precision. The results of the miRWalk algorithm and other are presented in Fig. 5. miRWalk algorithm successfully obtained (97.93% Accuracy, 98.88% Recall and 98.98% Precision) on the input gene sets and the results identified by 8 other programs are shown in Fig. 5.

3.4. Overview of validated data among miRWalk, TarBase, miR2Disease, miRecords, and miRTarBase databases

We compared Validated Target module of miRWalk with TarBase, miR2Disease [44], miRecords, PhenomiR [45] and miRTarBase databases in terms of relationships documented on miRNAs,

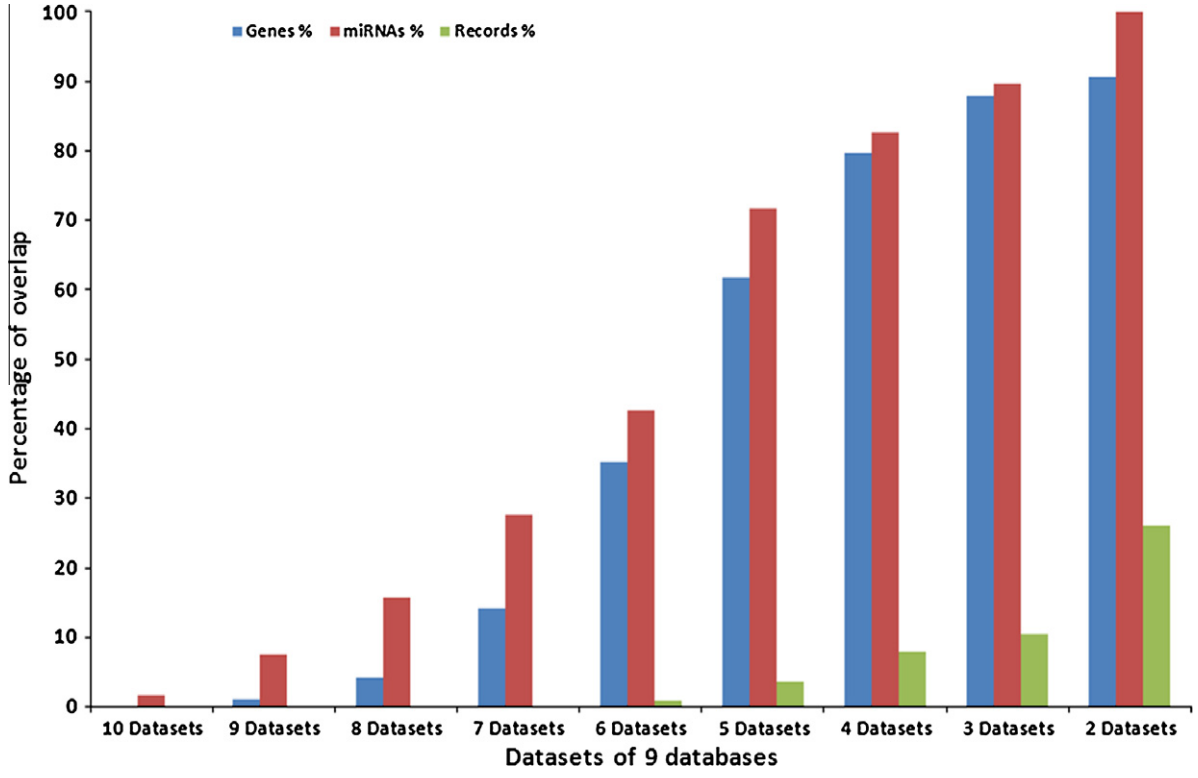

Fig. 4. Percentage of overlapping miRNA predictions across two or more databases. In comprehensive analysis, we compared predicted miRNAs binding sites in different combinations of 10 datasets and obtained a small overlap i.e. only 79 predictions (73 genes and 34 miRNAs) were common in all 10 datasets of 9 databases (DianamT, miRanda, miRDB, miRWalk, Pictar4/5, PITA, RNA22, RNAhybrid, Targetscan) considered. However, there is a considerable overlap between at least five datasets. See Supplementary Table 3 for more information.

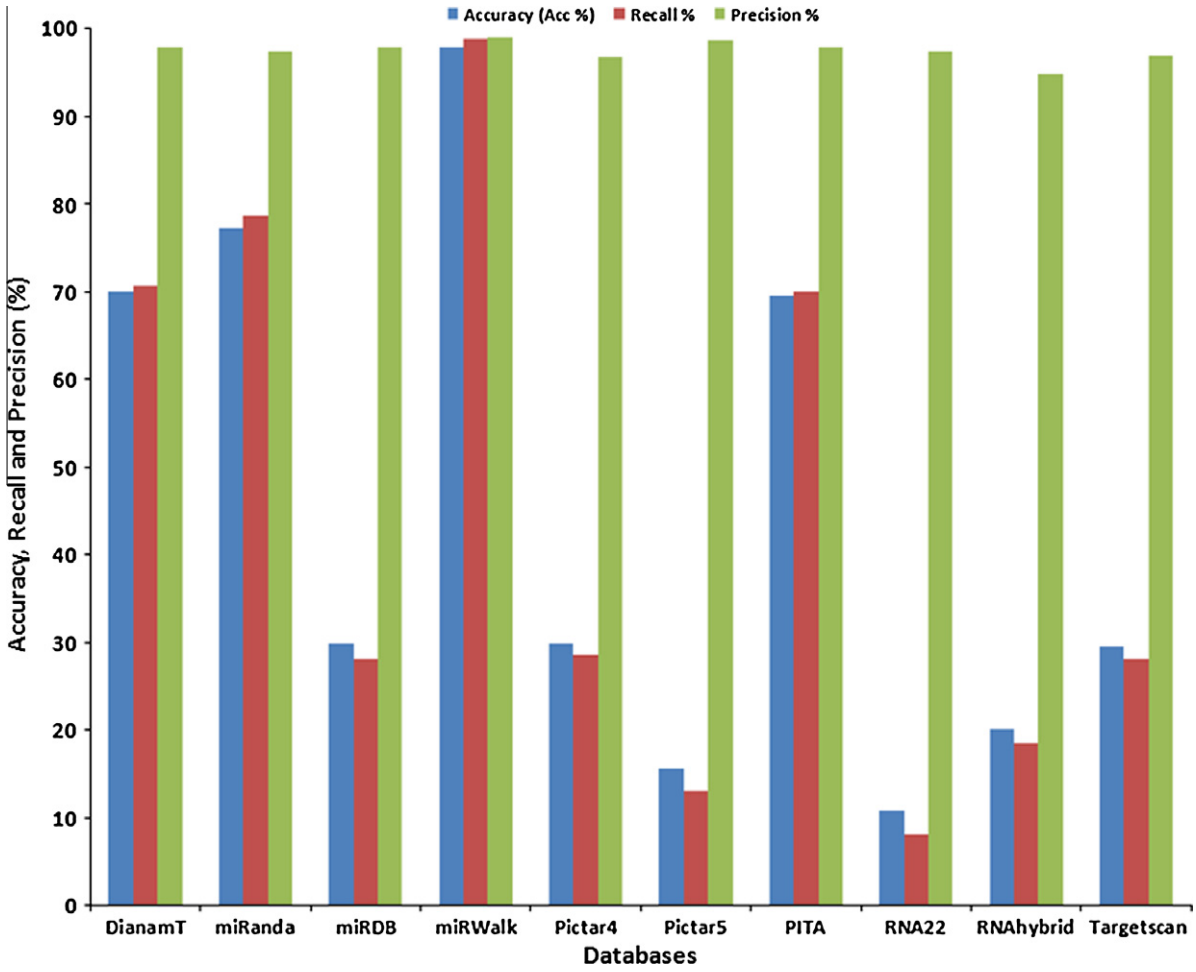

Fig. 5. Evaluation of miRWalk algorithm and eight established prediction programs. In this analysis, 1870 positive and 61 negative miRNA-target pairs were chosen to calculate the performance of the miRWalk algorithm and eight established prediction programs. In order to evaluate the performance of these different prediction programs, we used the statistical parameters, *viz*., Accuracy (Acc), Recall and Precision. These parameters are based on number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) and are calculated by using the following equations: Accuracy (Acc %) = (TP + TN)/(TP + TN + FP + FN) $*$ 100, Recall (%) = TP/(TP + FN) $*$ 100 and Precision (%) = TP/(TP + FP) $*$ 100.

genes, diseases, organs, cell lines, OMIM disorders, pathways, miRNA processing proteins (MPP) and PubMed articles. These databases are not only chosen due to their popularity, but also for surveying pertinent literature. For comprehensive analysis, we downloaded TarBase 5.0, miR2Disease (June 2010 release), miRecords (release 3.0) and miRTarBase (release 2.1) files from their respective websites and parsed the information only on human, mouse, and rat genes along with miRNAs. The validated information of PhenomiR was taken from the database home page and its publication. The Validated Target module of miRWalk database increased the validated information more than 3-fold (1572 miRNAs linked to 5080 genes, 691 diseases, 556 organs, 121 cell lines, 2033 OMIM disorders, 375 pathways and 70 MPP) as compared to above mentioned databases. Fig. 6 depicts an overview of the data stored in miRWalk and five other existing databases.

3.5. Evaluation of validated information of miRWalk database

For the evaluation process, a recall of EntrezGene PubMed articles was conducted. In details, first, we collected Entrez Gene Identifiers of human, mouse and rat miRNAs. Collected identifiers were compiled in a unique identifiers list. Second, a script was written in Perl programming language which automatically extracts all PubMed article Identifiers (PMID) which are marked as relevant for an input miRNA Entrez GeneID. A total of 1360 unique PMID were found to be marked as relevant. Afterwards, these PMIDs were queried against PubMed database to download their titles and abstracts. Then, the downloaded articles (only titles and abstracts) were scanned for miRNA name by using automated text-mining search module. Twelve hundred twenty-five out of 1360 abstracts were found to have at least one miRNA name present in their titles and/or abstracts. Thus, we obtained a recall of 90.07%. For the remaining 10% of articles, we randomly selected 50 PMIDs out of 135 and read their titles/abstracts for miRNA names. After reading, we observed that the randomly selected articles do not have any miRNA names present in their titles/abstracts, but other keywords like 'microRNAs', 'miRNAs', 'miRs', 'human genome sequencing' etc. were present. Since, several articles do not have any miRNA name in their titles/abstract. Therefore, we can expect that the automated text-mining search module can achieve a recall of more than 90.07%.

4. Discussion

miRWalk database is different from existing miRNA resources as: (i) a newly developed algorithm is used to predict all the possible miRNA binding sites by "walking" on the genes of three genomes (i.e. all protein coding genes, and their 10 kb upstream flanking regions and mitochondrial genes); (ii) the results of miRWalk are presented together with the results obtained from eight

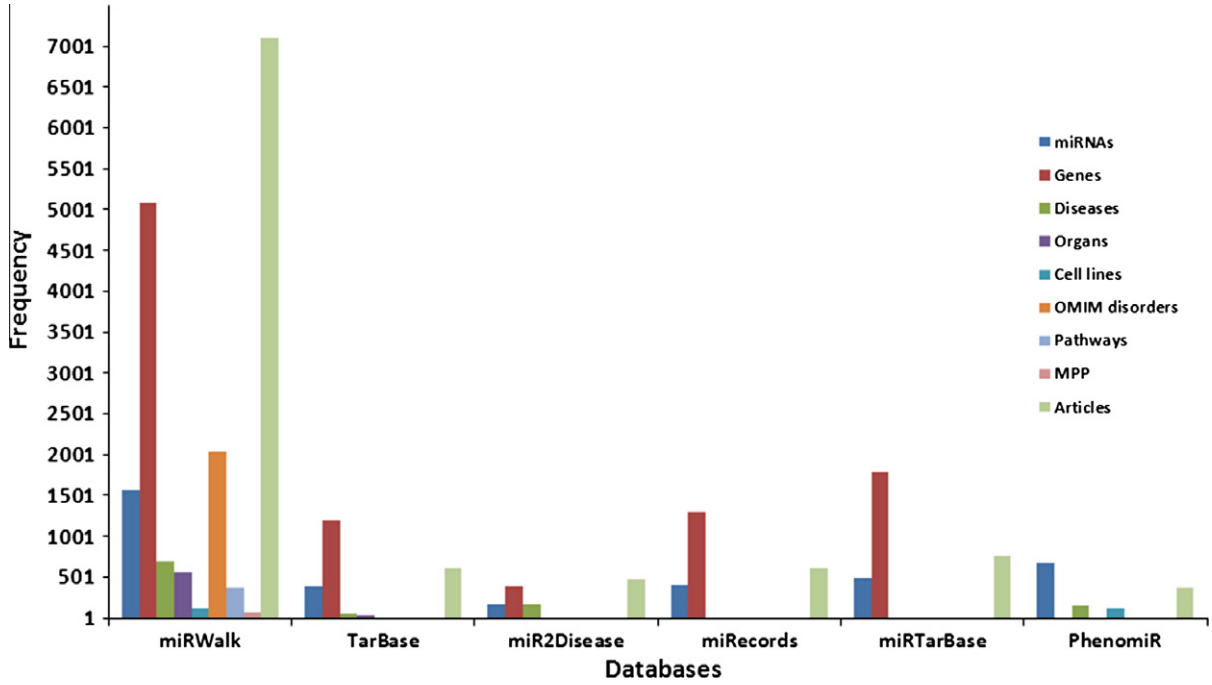

Fig. 6. Overview of validated data stored in miRWalk and five existing databases. Overview of the relationship information stored under Validated Target module of miRWalk and five existing database on miRNAs linked to genes, diseases, organs, OMIM disorders, cell lines, microRNA processing proteins and genes linked to human biological pathways. See Supplementary Table 4 for more information.

already established miRNA-target prediction programs for a comprehensive view of predicted miRNA binding sites on mRNA 3′-UTR region; (iii) miRWalk provides a more holistic view of genetic networks of miRNA-gene-pathways and miRNA-gene-OMIM disorder interactions; and (iv) miRWalk hosts new and unique features on experimentally validated miRNAs. Besides validated information, it also offers the information on proteins known to be involved in miRNA processing and available literature on miRNAs. A comparative analysis of features of miRWalk and already established databases on miRNAs is shown in Supplementary Table 5.

The 'Validated Target module' is updated every month and the 'Predicted Target module' is updated 6 months by executing automated Perl (www.perl.org) and Bioperl (www.bioper.org) scripts on the server of bwGRID Cluster (http://www.urz.uni-heidelberg.de/server/grid/index.en.html).

Similar to miRGator, which integrates three prediction databases (miRanda, PicTar and Targetscan), miRWalk integrates eight databases for a comprehensive study of predictions obtained from different algorithms. This allows the user to take more control over the prediction data that they consider. Not only does our resource conveniently incorporate eight different databases at one place, it also allows users to choose which combinations of databases they would like to consider for their search.

Some research groups have been adopting a new approach for the identification of new targets for known miRNAs. In this approach, miRNA expression in healthy and/or diseased tissue and/or organ was profiled by using miRNA microarray and statistical significant miRNAs were selected for further validation by Northern blot and/or q-PCR experiments. Afterwards, miRNA-target prediction programs were used to identify the possible target genes of these miRNAs. Finally, cell lines and/or animals were used for the knockdown of these miRNAs to measure the expression level of predicted genes. miRWalk database is helpful for this kind of approach, as a user can retrieve information on possible miRNA binding sites on the complete sequence or specific region(s) of targets by supplying miRNA names or uploading a file under Predicted Target module.

Moreover, the basic information on miRNAs (like mature, and stem loop sequence, identifiers, chromosome, strand and band) as well as other necessary data required for a miRNA research, such as regulatory binding sites on upstream and/or downstream flanking regions of pre-miRNA, information on the host gene of miRNA, which miRNAs share a similar seed with the user input miRNAs can be easily obtained. Furthermore, links are given, which redirect the user to NCBI, UCSC (for more annotations and information on miRNAs) and PubMed database for the quick and convenient access to available literature on the user input miRNAs associated with expression, diseases, organs and cell lines.

In the past couple of years, research groups demonstrated the involvement of miRNAs in complicated biological processes/pathways, including the control of developmental timing, haematopoietic cell differentiation, apoptosis, cell proliferation, organ development [10,11,46,47], as well as cancerogenesis [48–51] and other human diseases [40,52–57]. miR2Disease and PhenomiR are well known databases which offer information on miRNA interactions involved in diseases and biological processes. miR2Disease hosts 2663 relationships linked with 347 miRNAs whereas PhenomiR offers 12,192 relationships on 675 miRNAs linked to 146 diseases. In comparison to these resources, miRWalk database hosts 98,887 relationships on 1572 miRNAs from human, mouse and rat linked to 691 diseases. Thus, miRWalk presents much more information on miRNAs linked to human diseases.

Cell lines have been established in life sciences as easy to manipulate model systems for the study of cellular processes. A number of cell lines have been used to investigate the expression of miRNAs under different conditions. Such information is important because of the therapeutic potential of miRNAs in various diseases. miRWalk documents information on 121 cell lines linked to miRNA investigations.

Several studies in mice have demonstrated an important role of Dicer in the generation of mature miRNAs [58,59]. In contrast to the linear miRNA processing pathway that was initially thought to be universal for the biogenesis of all mature miRNAs, multiple discoveries led to the recognition of miRNA-specific differences,

that open a plethora of regulatory options to express and process individual miRNAs differentially. For instance, Drosha-mediated cleavage can regulate individual miRNAs: the hnRNP A1 binds specifically to pri-miR-18a and facilitates its processing [60]. Thus, information on proteins known to be involved in miRNA processing is helpful for researchers investigating the role of miRNAs in organ development and functional maintenance. This kind of information is lacking in the established miRNA databases hosting validated miRNA-target data, while it is implemented under the Validated Target module of miRWalk.

miRBase, TarBase, miRecords and miR2Disease are well-known databases for experimentally validated miRNA targets associated with genes, diseases, organs, and pathways. The Validated Targets module of miRWalk offers all the scattered information about miRNA interactions in a structured and uniform format. All the information of this module is extracted from PubMed database. Each of these entries is hyper-linked to PubMed database by using PubMed identifiers and allowing the retrieval of abstracts.

5. Conclusion

In conclusion, we developed a new platform on predicted as well as validated binding sites of miRNAs on the sequence of all known protein coding and mitochondrial genes of human, mouse and rat. In the future we will incorporate the same information on other species into miRWalk database. Furthermore the web interface will be improved and new modules for additional annotations will be added. Also an online tool for a motif search will be incorporated to query against input sequences of interest.

Acknowledgments

We are grateful of bwGRID Cluster Heidelberg (high performance cluster) for allowing us to use their supercomputing facility for the faster identification, and calculation of miRNA data. We thank Mr. Harald Schoppmann for his continuous help in the implementation of miRWalk database under a new server. We also wish to thank Sabine Neudecker and Asawari Kharkar for their help in improving the manuscript. This work was funded by the Research Council through Graduiertenkolleg 886 and by the German Federal Ministry of Research and Education through the National Genome Research Network (NGFN-2, Grant no. 01GR0450).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2011.05.002.

References

[1] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;116:281–97.
[2] Li LC, Okino ST, Zhao H, Pookot D, Place RF, Urakami S, et al. Small dsRNAs induce transcriptional activation in human cells. Proc Natl Acad Sci USA 2006;103:17337–42.
[3] Cullen BR. Transcription and processing of human microRNA precursors. Mol Cell 2004;16:861–5.
[4] Lai EC. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. Nat Genet 2002;30:363–4.
[5] Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell 2003;115:787–98.
[6] Doench JG, Sharp PA. Specificity of microRNA target selection in translational repression. Genes Dev 2004;18:504–11.
[7] Meister G. MiRNAs get an early start on translational silencing. Cell 2007;131:25–8.
[8] Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 2008;9:102–14.
[9] Griffiths-Jones S. MiRBase: the microRNA sequence database. Methods Mol Biol 2006;342:129–38.
[10] Chang S, Johnston Jr RJ, Frokjaer-Jensen C, Lockery S, Hobert O. MicroRNAs act sequentially and asymmetrically to control chemosensory laterality in the nematode. Nature 2004;430:785–9.
[11] Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, et al. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 2000;403:901–6.
[12] Yang N, Coukos G, Zhang L. MicroRNA epigenetic alterations in human cancer: one step forward in diagnosis and treatment. Int J Cancer 2008;122:963–8.
[13] Yu SL, Chen HY, Chang GC, Chen CY, Chen HW, Singh S, et al. MicroRNA signature predicts survival and relapse in lung cancer. Cancer Cell 2008;13:48–57.
[14] Doran J, Strauss WM. Bio-informatic trends for the determination of miRNA-target interactions in mammals. DNA Cell Biol 2007;26:353–60.
[15] Jones-Rhoades MW, Bartel DP. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell 2004;14:787–99.
[16] Shahi P, Loukianiouk S, Bohne-Lang A, Kenzelmann M, Kuffer S, Maertens S, et al. Argonaute – a database for gene regulation by mammalian microRNAs. Nucleic Acids Res 2006;34:D115–8.
[17] Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 2010;11:R90.
[18] Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. Nat Genet 2005;37:495–500.
[19] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005;120:15–20.
[20] Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. RNA 2004;10:1507–17.
[21] Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. MiRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 2009;37:D105–10.
[22] Nam S, Kim B, Shin S, Lee S. MiRGator: an integrated system for functional annotation of microRNAs. Nucleic Acids Res 2008;36:D159–64.
[23] Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, et al. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell 2006;126:1203–17.
[24] Wang X, El Naqa IM. Prediction of both conserved and nonconserved microRNA targets in animals. Bioinformatics 2008;24:325–32.
[25] Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet 2007;39:1278–84.
[26] Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, et al. DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res 2009;37:W273–6.
[27] Tay Y, Zhang J, Thomson AM, Lim B, Rigoutsos I. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature 2008;455:1124–8.
[28] Guang S, Bochner AF, Pavelec DM, Burkhart KB, Harding S, Lachowiec J, et al. An Argonaute transports siRNAs from the cytoplasm to the nucleus. Science 2008;321:537–41.
[29] Place RF, Li LC, Pookot D, Noonan EJ, Dahiya R. MicroRNA-373 induces expression of genes with complementary promoter sequences. Proc Natl Acad Sci USA 2008;105:1608–13.
[30] Lytle JR, Yario TA, Steitz JA. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. Proc Natl Acad Sci USA 2007;104:9667–72.
[31] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res 2000;28:15–8.
[32] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, et al. Ensembl 2011. Nucleic Acids Res 2011;39:D800–6.
[33] Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. MiRBase: tools for microRNA genomics. Nucleic Acids Res 2008;36:D154–8.
[34] Fundel K, Guttler D, Zimmer R, Apostolakis J. A simple approach for protein name identification: prospects and limits. BMC Bioinformatics 2005;6(Suppl. 1):S15.
[35] Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). Hum Mutat 2000;15:57–61.
[36] Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res 2002;30:42–6.
[37] Rajewsky N. MicroRNA target predictions in animals. Nat Genet 2006;38(Suppl):S8–S13.
[38] Sethupathy P, Megraw M, Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. Nat Methods 2006;3:881–6.
[39] Megraw M, Sethupathy P, Corda B, Hatzigeorgiou AG. MiRGen: a database for the study of animal microRNA genomic organization and function. Nucleic Acids Res 2007;35:D149–55.
[40] Pandey P, Brors B, Srivastava PK, Bott A, Boehn SN, Groene HJ, et al. Microarray-based approach identifies microRNAs and their target functional patterns in polycystic kidney disease. BMC Genomics 2008;9:624.
[41] Shah PK, Perez-Iratxeta C, Bork P, Andrade MA. Information extraction from full text scientific articles: where are the keywords? BMC Bioinformatics 2003;4:20.
[42] Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG. The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Res 2009;37:D155–8.

[43] Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, et al. MiRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res 2011;39:D163–9.
[44] Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, et al. MiR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 2009;37:D98–D104.
[45] Ruepp A, Kowarsch A, Schmidl D, Bruggenthin F, Brauner B, Dunger I et al. PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. Genome Biol 11: R6.
[46] Johnston RJ, Hobert O. A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. Nature 2003;426:845–9.
[47] Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993;75:843–54.
[48] Gregory RI, Shiekhattar R. MicroRNA biogenesis and cancer. Cancer Res 2005;65:3509–12.
[49] He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, et al. A microRNA polycistron as a potential human oncogene. Nature 2005;435:828–33.
[50] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature 2005;435:834–8.
[51] McManus MT. MicroRNAs and cancer. Semin Cancer Biol 2003;13:253–8.
[52] Blenkiron C, Goldstein LD, Thorne NP, Spiteri I, Chin SF, Dunning MJ, et al. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. Genome Biol 2007;8:R214.
[53] Calin GA, Croce CM. MicroRNA-cancer connection: the beginning of a new tale. Cancer Res 2006;66:7390–4.
[54] Calin GA, Sevignani C, Dumitru CD, Hyslop T, Noch E, Yendamuri S, et al. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proc Natl Acad Sci USA 2004;101:2999–3004.
[55] Eisenberg I, Eran A, Nishino I, Moggio M, Lamperti C, Amato AA, et al. Distinctive patterns of microRNA expression in primary muscular disorders. Proc Natl Acad Sci USA 2007;104:17016–21.
[56] Lee SO, Masyuk T, Splinter P, Banales JM, Masyuk A, Stroope A, et al. MicroRNA15a modulates expression of the cell-cycle regulator Cdc25A and affects hepatic cystogenesis in a rat model of polycystic kidney disease. J Clin Invest 2008;118:3714–24.
[57] McCarthy JJ, Esser KA, Andrade FH. MicroRNA-206 is overexpressed in the diaphragm but not the hindlimb muscle of mdx mouse. Am J Physiol Cell Physiol 2007;293:C451–7.
[58] Ho J, Ng KH, Rosen S, Dostal A, Gregory RI, Kreidberg JA. Podocyte-specific loss of functional microRNAs leads to rapid glomerular and tubular injury. J Am Soc Nephrol 2008;19:2069–75.
[59] Shi S, Yu L, Chiu C, Sun Y, Chen J, Khitrov G, et al. Podocyte-selective deletion of dicer induces proteinuria and glomerulosclerosis. J Am Soc Nephrol 2008;19:2159–69.
[60] Guil S, Caceres JF. The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. Nat Struct Mol Biol 2007;14:591–6.