

# miRvestigator: web application to identify miRNAs responsible for co-regulated gene expression patterns discovered through transcriptome profiling

Christopher L. Plaisier, J. Christopher Bare and Nitin S. Baliga\*

Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109-5234, USA

Received February 17, 2011; Revised April 22, 2011; Accepted April 29, 2011

## ABSTRACT

Transcriptome profiling studies have produced staggering numbers of gene co-expression signatures for a variety of biological systems. A significant fraction of these signatures will be partially or fully explained by miRNA-mediated targeted transcript degradation. miRvestigator takes as input lists of co-expressed genes from *Caenorhabditis elegans*, *Drosophila melanogaster*, *G. gallus*, *Homo sapiens*, *Mus musculus* or *Rattus norvegicus* and identifies the specific miRNAs that are likely to bind to 3' un-translated region (UTR) sequences to mediate the observed co-regulation. The novelty of our approach is the miRvestigator hidden Markov model (HMM) algorithm which systematically computes a similarity *P*-value for each unique miRNA seed sequence from the miRNA database miRBase to an overrepresented sequence motif identified within the 3'-UTR of the query genes. We have made this miRNA discovery tool accessible to the community by integrating our HMM algorithm with a proven algorithm for *de novo* discovery of miRNA seed sequences and wrapping these algorithms into a user-friendly interface. Additionally, the miRvestigator web server also produces a list of putative miRNA binding sites within 3'-UTRs of the query transcripts to facilitate the design of validation experiments. The miRvestigator is freely available at <http://mirvestigator.systemsbiology.net>.

## INTRODUCTION

MiRNAs post-transcriptionally regulate expression of protein coding genes by selectively recruiting protein complexes to un-translated regions (UTRs) of target transcripts. The miRNA effector complex primarily recognizes the target transcripts through sequence complementarity

to the 8 bp miRNA seed region that has been discovered to be preferentially located in 3'-UTRs (1,2). Thus by searching the 3'-UTRs of putatively co-regulated genes for an overrepresented sequence that corresponds to a miRNA seed region, it is feasible to discover the identity of the regulatory miRNA (3). A recent meta-analysis of co-expression signatures from 46 different cancers identified miRNAs using this approach (3). While this provides proof-of-concept for *de novo* discovery of miRNAs simply from knowledge of which genes are co-expressed, the sensitivity of the algorithm itself was lacking and able to make miRNA calls for only four of over 2000 co-expression signatures.

We constructed the miRvestigator to provide users a fully functional web server that reverse engineers miRNA mediated gene regulation in a flexible, intuitive and user-friendly interface. The novelty of our approach is an algorithm that compares an overrepresented sequence motif as a matrix to a miRNA seed sequence as a string to systematically identify the most likely complementary miRNA seed. This yields a more probabilistic solution that does not require any heuristics to convert the overrepresented sequence motif into a consensus string. The miRvestigator framework is designed to identify miRNA mediated regulation which is one of several regulators of gene expression including transcription factors. The miRvestigator web server takes as input a set of genes (Entrez gene, Ensembl gene, RefSeq transcript IDs or official gene symbol) from one of six different species: *Caenorhabditis elegans*, *Drosophila melanogaster*, *G. gallus*, *Homo sapiens*, *Mus musculus* or *Rattus norvegicus*.

There are three steps to the discovery of miRNA-mediated gene regulation: (i) discover co-expressed genes; (ii) discover conserved cis-regulatory signature in UTRs of co-expressed genes as evidence of miRNA-mediated control; and finally (iii) identify the miRNA seed sequence from miRBase that is complementary to the cis-regulatory signature (4). With regard to the first step, our goal was to allow users the flexibility to use any method to select a gene set although it is expected

\*To whom correspondence should be addressed. Tel: +1 206 732 1200; Fax: +1 206 732 1299; Email: [nbaliga@systemsbiology.org](mailto:nbaliga@systemsbiology.org)

that these genes are likely to be co-regulated by a common factor. A frequently used approach applies clustering of transcriptional profiles as a means to identify co-expressed genes. Biclustering approaches such as cMonkey (5) are helpful in addressing this issue by allowing the discovery of conditionally co-expressed genes. For the second step, a number of miRNA target site prediction algorithms have been published (6–9). However, a potential issue of such methodology is the flawed assumption that activity of a miRNA is constitutive and observable across a wide array of environmental conditions. SylArray (10) identifies a putative miRNA-binding site using the 3'-UTR word enrichment Sylamer algorithm (11). In contrast, the miRvestigator uses a *de novo* motif discovery algorithm that models miRNA binding in a probabilistic manner. Unlike the Sylamer algorithm which works on a sorted gene list, miRvestigator expects a specifically selected subset of co-expressed genes identified using classification methods (e.g. hierarchical clustering, biclustering, etc.).

miRvestigator scans the 3'-UTR sequences of query genes for an overrepresented sequence motif using the Weeder software package (12–14). However, miRvestigator has been designed to be modular so a user can exchange or supplement Weeder with a different motif detection algorithm. After detecting an overrepresented sequence motif, miRvestigator then attempts to identify the cognate miRNA regulator of the co-expressed genes. We accomplished this by developing the miRvestigator hidden Markov model (HMM) to systematically compare an overrepresented sequence motif which is typically in the form of a position specific scoring matrix (PSSM) to a database of miRNA seed sequences (strings) from miRBase. miRvestigator converts the overrepresented PSSM into a profile HMM and scores the complementarity of each unique miRNA seed sequence against this model using the Viterbi algorithm. The Viterbi algorithm simultaneously aligns and computes the probability of base pair complementarity between the miRNA seed sequence and an overrepresented PSSM from 3'-UTR sequences. The significance of the complementarity probability for a given unique miRNA seed sequence is calculated by exhaustively computing the complete distribution of complementarity probabilities for all potential *k*-mer seed sequences (where *k* = 6, 7 or 8-mer). In this manner, starting with sets of genes with evidence for co-expression a user is able to reverse engineer miRNA regulation by using a tandem of *de novo* motif detection and systematic identification of complementary miRNA seeds by miRvestigator.

#### Using miRvestigator input gene list format

Currently miRvestigator is able to accept as input a list of Entrez gene, Ensembl gene, RefSeq transcript IDs or official gene symbols. We recommend converting any other type of identifier to Entrez gene IDs using the DAVID bioinformatics resource Gene ID Conversion Tool (<http://david.abcc.ncifcrf.gov/conversion.jsp>) (15). The input gene list should be separated by comma, space, newline, or tab delimiters.

#### Parameters

Users must first select the source of genes from one of six different species: *C. elegans*, *D. melanogaster*, *G. gallus*, *H. sapiens*, *M. musculus* or *R. norvegicus*. Next, they specify parameters for *de novo* motif detection by Weeder: (i) motif length can be set to either 6 or 8 bp; (ii) either the default background model for the species in Weeder or a 3'-UTR specific background model is selected; (iii) select a quality threshold for the predicted sites in the 3'-UTRs of co-regulated genes based on identity to the overrepresented motif. The users will then choose the parameters for the miRvestigator HMM by first choosing the miRNA seed models they wish to include in a run (i.e. 6, 7 or 8-mer models). If the overrepresented PSSM is shorter than 8 bp, the models are constrained to the length of the PSSM. The user can also decide to model wobble base pairing and can specify a minimum frequency for the G or U nucleotides before a wobble base pairing is considered, we recommend a frequency greater than or equal to random (25% G or U). Finally the user can specify the number of resulting complementary miRNA seeds returned, and an optional e-mail address so that they can be notified with a link to the results.

The default parameters for miRvestigator search for only 8-bp motifs using the default Weeder background model and compare to all miRNA seed models (6, 7 and 8-mer) without modeling wobble base pairing should perform adequately for typical analysis. The default background model comes packaged with Weeder and is built based upon the sequences upstream of the start codon of all annotated genes. This provides contrast for identifying miRNA seed motifs in 3'-UTRs, but in certain cases we have found a background model constructed from the annotated 3'-UTR sequences to perform better. We caution against relying on 6-bp overlaps between a motif and a miRNA from miRBase as it is very likely that this may happen by chance. Also modifying the target-site quality to values <95% leads to identification of binding sites with >1 bp difference relative to the motif consensus. Both 6bp and more degenerate target sites may be useful and were added for hypothesis generation; however, their validity must be experimentally evaluated.

#### miRvestigator analysis workflow

**3'-UTR sequence acquisition.** After a user submits a query, miRvestigator retrieves 3'-UTR sequences for the set of input genes from a pre-computed database. The 3'-UTR sequence database was acquired by downloading the sequence and RefSeq gene definition files for *C. elegans*, *D. melanogaster*, *G. gallus*, *H. sapiens*, *M. musculus* and *R. norvegicus* from the UCSC genome browser FTP site (16). To minimize overlap the set of RefSeq genes that mapped to an Entrez gene were collapsed and the regulatory regions were merged to include all potential regulatory sequences. All annotated introns were removed as they are present only transiently in expressed transcripts. The RefSeq to Entrez gene mapping was downloaded from NCBI Gene FTP site (<ftp://ftp.ncbi.nih.gov/gene/DATA/gene2refseq.gz>). The 3'-UTR for genes without an

annotated 3'-UTR were estimated to be a sequence of length equal to the median of annotated 3'-UTRs downstream of the stop codon (9). The sequences for the input gene list are written to a FASTA file for *de novo* motif detection with Weeder and the number of genes successfully annotated with sequences is available on the status and results pages in the submission parameters section.

**Identifying overrepresented sequence motifs.** Weeder then analyzes the 3'-UTR sequences from the FASTA file created above to identify either 6 and/or 8 bp motifs (a parameter that can be specified) that correspond to the miRNA seed sequence. The Weeder algorithm is provided with a background model and returns a PSSM of the overrepresented sequence motif. Weeder also provides a list of predicted target sites of this oligo with a maximum of 2 bp of discrepancy in the 3'-UTRs of the miRNA co-regulated gene set. The predicted target sites are then filtered as having a similarity percent greater than or equal to the target site quality threshold (a parameter that can be specified). The duplex free energy of the putative target site and the reverse-complement of the motif consensus are calculated using the ViennaRNA RNAduplex program. The results of the Weeder analysis are entered into the results database and are accessed by the miRvestigator HMM to identify the most likely cognate miRNA.

**miRNA seed sequence acquisition.** The miRNA seed sequences for each species were downloaded in FASTA format as mature miRNA sequences via FTP from the miRBase miRNA database (<ftp://mirbase.org/pub/mirbase/CURRENT/mature.fa.gz>) (4). A filter is applied to identify the species of interest, the seed sequence (1–8 bp) is extracted for each miRNA in the database, and the seed sequences are then condensed into the unique set of seed sequences (for these studies 1064 unique seed sequences were identified from miRBase Release 16). The reverse-complement of the unique seed sequences are then used for subsequent comparison to the miRvestigator HMM. miRvestigator allows different models for the base-pairing between the miRNA seed to the cognate protein coding transcript binding sites (a parameter that can be specified as 6, 7 and/or 8-mer) (1,2). The models are provided to miRvestigator as trimmed miRNA seeds that are used as input into the miRvestigator HMM.

**Identifying the cognate miRNA for an overrepresented sequence motif.** The most likely cognate miRNA for each overrepresented motif discovered by Weeder is then computed using the miRvestigator HMM. First, the overrepresented motif PSSM is converted into a miRvestigator HMM. Then for each seed model all possible  $k$ -mers (where  $k = 6, 7$  or 8-mer) are exhaustively scored against the miRvestigator HMM to produce the distribution of Viterbi complementarity probabilities. In succession, each unique miRNA seed sequence from miRBase for the species of interest is scored against the miRvestigator HMM and the Viterbi complementarity

probability is compared to the exhaustive distribution to produce a Viterbi  $P$ -value. The results are entered into the results database and are retrieved for display on the results page.

**Retrieving results.** Results can be retrieved through three different mechanisms:

- (i) Users are notified by email upon completion of miRvestigator analysis.
- (ii) Users can check status or retrieve results by clicking on a bookmark.
- (iii) Results can be retrieved with a unique job identifier.

### miRvestigator implementation

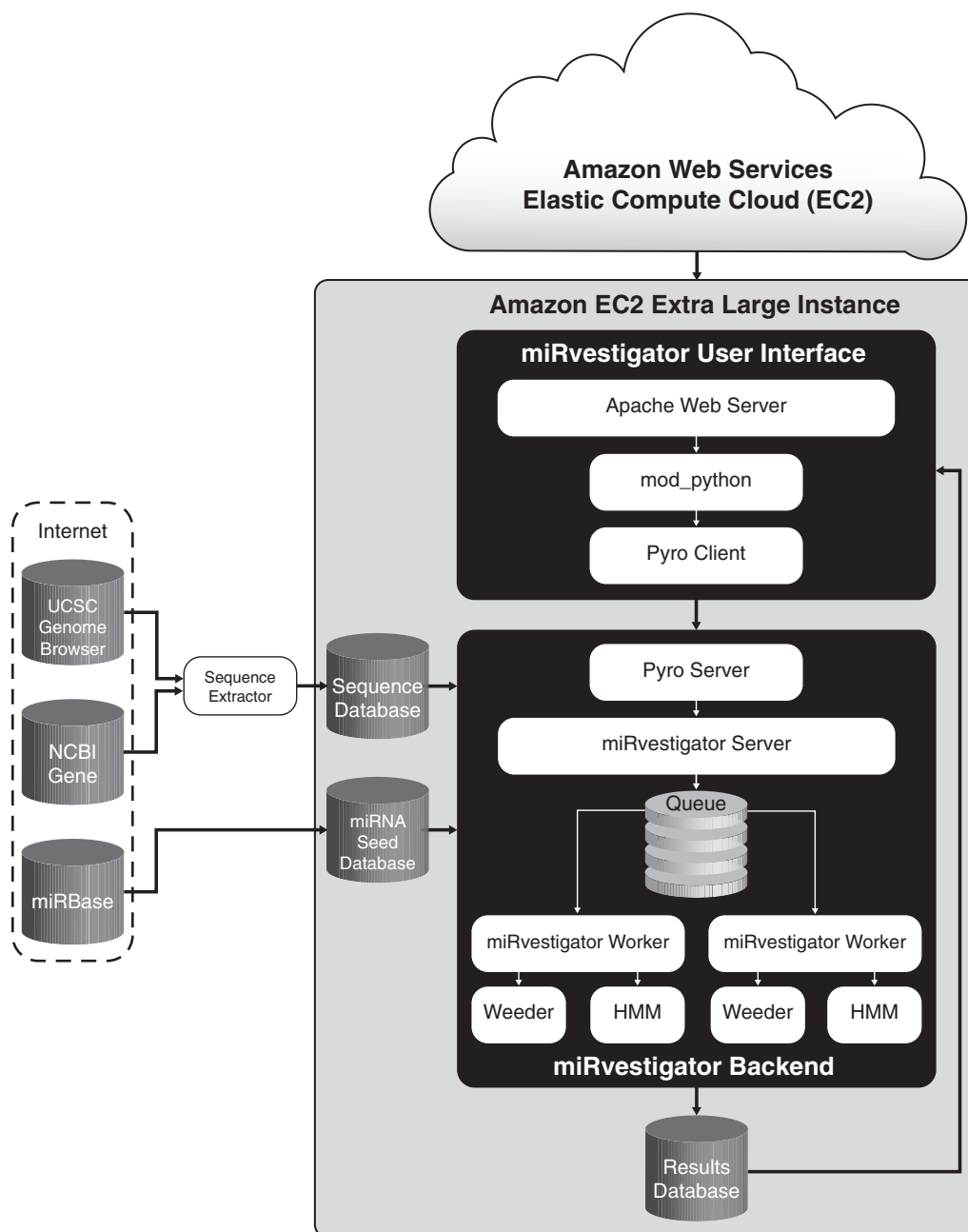
miRvestigator is an open source web application built with Python (v2.6), MySQL, and Apache (Figure 1). Code is available on GitHub ([http://github.com/cplaisier/miRvestigator\\_www](http://github.com/cplaisier/miRvestigator_www)) and a stand-alone application for use with motifs is also available ([http://github.com/cplaisier/miRvestigator\\_standalone](http://github.com/cplaisier/miRvestigator_standalone)). The miRvestigator has been designed to accommodate the sizable amount of processing required by the algorithm, the majority of which goes into computing the background distribution of Viterbi probabilities. The program is organized into a web tier and a backend process that manages several child worker processes. The web tier is hosted within an Apache web server with mod\_python. Upon user request, the web tier packages request parameters and pass them to the backend for execution via Pyro connection. The miRvestigator backend receives requests from the user interface and pushes them onto a task queue. The next available worker pulls the job off the queue, executes the miRvestigator analysis pipeline (Weeder and the miRvestigator HMM) using sequence data for the appropriate organism and the miRNA seed database and records the results. Finally, the web tier accesses and displays the results.

The application is hosted on the Amazon Web Services Elastic Compute Cloud (EC2), providing scalability to meet future demand. The miRvestigator machine image is available for users who wish to deploy customized or private instances. Currently, miRvestigator runs on an Extra Large Instance with 15GB of memory, 8 EC2 Compute Units, 1690GB of local instance storage on a 64-bit Ubuntu Linux operating system. The current compute time for 13 genes is 2min and for 106 genes is 10.8 min.

miRvestigator relies upon a number of Python libraries including multiprocessing, and Pyro. Sequence logo plots as Portable Network Graphics (PNG) images are generated using the WebLogoLib python library (<http://code.google.com/p/weblogo/>).

### miRvestigator use case and output description

We demonstrate the utility of miRvestigator by examining gene expression changes resulting from intentional perturbation of a known miRNA in mouse. In brief, 372 genes were up-regulated upon intravenous administration of mmu-miR-122 (MIMAT0000246) antagomir (17).

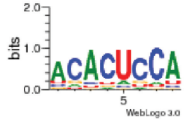


**Figure 1.** This implementation diagram describes the miRvestigator web server currently residing on the Amazon Web Services Elastic Compute Cloud (EC2) as an Extra Large Instance. The miRvestigator is organized into two levels, a web tier for user interaction and the backend where the bulk of computation occurs. Results are stored in a MySQL database which also contains the 3'-UTR sequence and miRNA seed sequences for each of the six species. The 3'-UTR sequences are generated by integrating information from the UCSC genome browser and NCBI Entrez gene database. The miRNA seed sequences are retrieved from miRBase.

Taking the list of 372 genes as input miRvestigator accurately identified mmu-miR-122 as the cognate miRNA with a perfectly complementary 8-mer base pairing (Figure 2A). This analysis recapitulated known biology by accurately demonstrating that inhibition of mmu-miR-122 was the primary cause for up-regulation of the 372 genes, of which 3'-UTRs of 11% of the genes were predicted to contain a perfect 8-mer binding site (95% target site quality threshold).

All of the associated results from such analysis can be accessed by a user to make an informed assessment and design appropriate validation experiments. For instance, in the use case described above, a table of the most likely cognate miRNAs that were discovered in this analysis can be retrieved by clicking on the logo plot of the Weeder motif (Figure 2B). Further, a table of predicted miRNA-binding sites is also generated and accessed by clicking on the percent genes that are predicted to possess matches to

**A** Summary of results

Summary of Results				
Motif	Top miRNA	Complementary Base-Pairing	Complementarity P-Value	% of Input Sequences with Site
	mmu-miR-122	Motif 5' ACACUCCA 3'       3' UGUGAGGU 5' miRNA Seed	1.5e-05	11%

**B** Complementary miRNAs

Top 10 miRNAs Complementary to the Weeder Motif ACACUCCA [+]					
<a href="#">Download table as CSV</a>					
miRNA Name	miRNA Seed	Seed Model	Length of Complementarity	Complementary Base-Pairing	Complementarity P-Value
mmu-miR-122	UGGAGUGU	8mer	8	Motif 5' ACACUCCA 3'       3' UGUGAGGU 5' miRNA Seed	1.5e-05
mmu-miR-1 mmu-miR-206	UGGAAUGU	8mer	8	Motif 5' ACACUCCA 3'       3' UGUAAGGU 5' miRNA Seed	2.0e-04
mmu-miR-3068	UGGAGU	6mer_2	6	Motif 5' ACACUCCA 3'       3' --UGAGGU 5' miRNA Seed	2.4e-04
mmu-miR-1	GGAGUG	6mer_2	6	Motif 5' ACACUCCA 3'       3' GUGAGGU 5' miRNA Seed	4.0e-04

**C** Putative miRNA binding sites

Position of Putative miRNA Binding Sites in Submitted Genes for the Weeder Motif ACACUCCA [+]					
<a href="#">Download table as CSV</a>					
Gene	Gene symbol	Sequence of Site	Start Relative to Stop Codon (bp)	% Similarity to Consensus Motif (Quality = High   Medium   Fair)	Minimum Free Energy (MFE) of mRNA-miRNA Duplex
18555	Cdk16	ACACUCCA	1028	100.00	-11.40
18393	Orc2	ACACUCCA	340	100.00	-11.40
235339	Dlat	ACACUCCA	1413	100.00	-11.40
67991	Nacc2	ACACUCCA	413	100.00	-11.40
67759	5033414D02Rik	ACACUCCA	172	100.00	-11.40

**Figure 2.** miRvestigator results for an *in vivo* study where 372 genes were up-regulated when miR-122 was inhibited by antagomir in mice. (A) Summary of the results for the run. The table contains a motif logo plot, the top miRNA(s) complementary to the motif, the base pairing for each miRNA seed to the motif, the Viterbi *P*-Value of the significance of the complementarity between the motif and the miRNA seed, and the percent of the input sequences with a predicted miRNA site. The summary table contains an entry for each motif size (6 or 8 bp). (B) Table of top miRNA

(continued)

the motif consensus equal to or greater than the quality threshold (Figure 2C). To facilitate further exploration of these results, the miRNA identifiers are linked to corresponding entries in miRBase and the official gene symbol or Entrez identifiers links to the NCBI Entrez Gene database. It is also possible to download either the most likely cognate miRNA or predicted miRNA target sites tables as a comma separated values (CSV) file.

### Future improvements

We are planning to integrate additional *de novo* motif search algorithms to potentially improve discovery of overrepresented motifs. We are also considering analyses that begin with user-specified motifs to identify matching miRNAs in miRBase in a manner similar to STAMP, a web tool for exploring transcription factor binding motifs (18).

### CONCLUSIONS

The miRvestigator web server provides biologists with a powerful suite of algorithms that together identify miRNAs that bind and regulate genes discovered to be co-expressed in transcriptome profiling studies. The website takes as input a list of gene identifiers (Entrez gene, Ensembl gene, RefSeq transcript or official gene symbol) for one of six different species (*C. elegans*, *D. melanogaster*, *G. gallus*, *H. sapiens*, *M. musculus* or *R. norvegicus*). For a given set of genes, miRvestigator: (i) extracts their relevant 3'-UTR sequences; (ii) scans these 3'-UTR sequences for an overrepresented motif using the Weeder algorithm; (iii) identifies putative binding sites for the overrepresented motif; and (iv) applies a HMM and the Viterbi algorithm to identify the miRNA(s) that is most likely to bind to the overrepresented motif. The user friendly web interface makes it easy for biologists to tune parameters and submit jobs to miRvestigator. The output from miRvestigator is a ranked list of miRNAs presented in tabular format with links to corresponding records in miRBase, statistical assessment of complementarity quality, and an alignment of the motif to the miRNA seed sequences. The output also includes a second table with the putative binding locations of the miRNA within the 3'-UTRs of query genes. The miRvestigator provides a valuable tool for users to identify miRNAs regulating biological processes of interest and the information required to design experiments to test these predictions.

### FUNDING

Funding for open access charge: National Institutes of Health (grants P50GM076547 and 1R01GM077398-01A2); DoE (DE-FG02-04ER64685); NSF (DBI-0640950); Luxembourg Centre for Systems Biomedicine; we thank the University of Luxembourg for support.

*Conflict of interest statement.* None declared.

### REFERENCES

- Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.
- Brennecke,J., Stark,A., Russell,R.B. and Cohen,S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, e85.
- Goodarzi,H., Elemento,O. and Tavazoie,S. (2009) Revealing global regulatory perturbations across human cancers. *Mol. Cell*, **36**, 900–911.
- Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
- Reiss,D.J., Baliga,N.S. and Bonneau,R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, **7**, 280.
- Megraw,M., Sethupathy,P., Corda,B. and Hatzigeorgiou,A.G. (2007) miRGen: a database for the study of animal microRNA genomic organization and function. *Nucleic Acids Res.*, **35**, D149–D155.
- Friedman,R.C., Farh,K.K., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.
- Betel,D., Wilson,M., Gabow,A., Marks,D.S. and Sander,C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res.*, **36**, D149–D153.
- Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
- Bartonicek,N. and Enright,A.J. (2010) SylArray: a web server for automated detection of miRNA effects from expression data. *Bioinformatics*, **26**, 2900–2901.
- van Dongen,S., Abreu-Goodger,C. and Enright,A.J. (2008) Detecting microRNA binding and siRNA off-target effects from expression data. *Nat. Methods*, **5**, 1023–1025.
- Pavesi,G., Mereghetti,P., Zambelli,F., Stefani,M., Mauri,G. and Pesole,G. (2006) MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Res.*, **34**, W566–W570.
- Fan,D., Bitterman,P.B. and Larsson,O. (2009) Regulatory element identification in subsets of transcripts: comparison and integration of current computational methods. *RNA*, **15**, 1469–1482.
- Linhart,C., Halperin,Y. and Shamir,R. (2008) Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res.*, **18**, 1180–1189.
- Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

### Figure 2. Continued

matches sorted by complementarity to the identified motif. The miRNA name(s) are listed for each unique miRNA seed sequence (each miRNA name is a link to the miRBase entry), the unique miRNA seed, the seed model, the complementary base pairing of the miRNA seed sequence to the motif and the Viterbi *P*-Value for the significance of the complementarity. (C) Table of predicted miRNA-binding sites from Weeder. The official gene symbols or Entrez gene identifiers are listed for each predicted binding site (which is a link to NCBI Entrez gene database entry), the unmapped identifier, the sequence of the site, the location of the start of the site relative to the stop codon, the similarity of the site to the identified motif (100% = perfect complementarity, 95% = 1-bp difference, etc.) and the duplexing free energy for the reverse complement of the motif consensus to the predicted target site.

16. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H., Coelho,A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
17. Krützfeldt,J., Rajewsky,N., Braich,R., Rajeev,K.G., Tuschl,T., Manoharan,M. and Stoffel,M. (2005) Silencing of microRNAs in vivo with 'antagomirs'. *Nature*, **438**, 685–689.
18. Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–W258.