High-throughput Computational Characterization and Prediction of MicroRNA Targets

by

Xiao Fan

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Signal and Image processing

Electrical and Computer Engineering
University of Alberta

Abstract

MicroRNAs (miRNAs) are short (~21 nucleotides) endogenous noncoding RNAs. They are widespread post-transcriptional regulators in eukaryotes that bind target messenger RNAs (mRNAs) and regulate the protein expression levels. MiRNAs have attracted substantial amount of research attention and consequently thanks to sequencing effort their counts continually increase over the past decade. We contributed to these efforts by designing, building, and applying a comprehensive platform for end-to-end processing of miRNA data generated by next generation sequencing. The platform, which integrates multiple computational tools, filters out known miRNAs, discovers new miRNAs, and quantifies differential expression among samples. The key element to decipher functional roles of the fast growing number of miRNAs is the high-throughput identification of miRNA targets. Computational prediction methods are widely used for this purpose. We review a comprehensive collection of 38 miRNA target predictors in animals that were developed over the last decade. Our in-depth analysis considers all significant perspectives including the underlying methodologies, ease of use, availability, impact, and evaluation protocols. We comparatively evaluate seven representative methods when predicting targets at different levels of annotations and when predicting different types of targets. As one of observations we found on average only 7% of non-canonical miRNA targets which have <7 Watson-Crick base pairs in the seed region (nucleotides 1–8 from 5' end of the miRNA) can be identified by current miRNA target predictors. Moreover, our large scale analysis of 3' UTR regions in several databases reveals that about half of miRNA targets are non-canonical. These targets are prevalent and hard to predict, which motivated us to develop the first custom-designed high-throughput method that accurately predicts the non-canonical targets solely from the miRNA and target sequences. Empirical tests on targets annotated with low-throughput

methods, microarrays, RNA-seq and pSILAC show that our method correctly predicts 40% of non-canonical targets and more accurately finds highly repressed genes when compared to the existing methods.

Preface

This thesis is an original work by Xiao Fan. The sequencing data and biochemical experiments in Chapter 3 were provided by Dr. Marek Michalak's group. Chapters 3 and 4 have been published. Dr. Lukasz Kurgan supervised this published work and contributed to the conceptualization of these projects and writing and revising the manuscripts.

Acknowledgements

First and foremost, I would like to express my deep gratitude to my supervisor Dr. Lukasz Kurgan for all his guidance, passion, motivation and most of all for the tremendous amount of time he spent to supervise my work and make me a better researcher. I could not have imagined having a better mentor for my Ph.D. study.

I would like to thank my parents and my husband for their unconditional love, understanding and support.

I would like to thank my fellow lab members for their collaboration and support.

I would like to thank my friends for their support, motivation and time spent together.

I would like to extend my gratitude to Alberta Innovates Technology Futures for their financial assistance during my Ph.D. study.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AUC: area under the ROC curve

DF: degree of freedom

DNA: deoxyribonucleic acid

ER: endoplasmic reticulum

FN: false negatives

FP: false positives

GEO: gene expression omnibus

GO: Gene Ontology

MCC: Matthews correlation coefficient

miRNA: microRNA

mRNA: messenger RNA

PNR: predicted-to-native positive rate

pSILAC: pulsed stable isotope labeling by/with amino acids in cell culture

qPCR: real-time polymerase chain reaction

RBF: radial basis function

RISC: RNA-induced silencing complex

RNA: ribonucleic acid

ROC: receiver operating characteristic

SNR: signal-to-noise ratio

SVM: support vector machine

TN: true negatives

TP: true positives

UTR: untranslated region

WC: Watson-Crick

Chapter 1

Introduction

MicroRNAs (miRNAs) are abundant and short endogenous noncoding RNAs composed of 19-23 nucleotides. The first lin-4 miRNA was characterized in *C. elegans* in 1993 [1], however, miRNAs were not recognized as a distinct class of biological regulators until the second one was discovered in 2000 [2]. This class of small RNAs is now known as widespread post-transcriptional regulators in eukaryotes. They regulate the expression levels of messenger RNA (mRNA) by binding (interacting with) the target mRNAs and ultimately regulating expression levels of the corresponding proteins. As indirect protein-expression regulators, miRNAs exert multiple cellular functions through proteins and recently attracted substantial amount of research attention. As a novel class of molecules, miRNAs also hold promise for medical breakthroughs in disease focused gene therapy. Consequently, the number of articles that are related to miRNAs has grown exponentially in recent years; currently almost 20,000 publications on this subject can be found in PubMed. The count of miRNA sequences has also registered a big growth in the last decade thanks to next generation sequencing efforts. Many pipelines have been developed to process miRNA sequencing data and discover new miRNAs [3-5]. To date, miRNAs have been used to study signal transduction and pathogenesis of genetic disorder including amyotrophic lateral sclerosis, Alzheimer's disease, Parkinson's disease, muscular dystrophies [6-9] and other disease states. They were already utilized in preclinical drug development, primarily as drug targets [10-14], and a few anti miRNAs have entered clinical trials [15]. Development of miRNA-directed novel therapeutics is under way [16-18] and miRNA-based targeting in cancer is not far behind [19-21]. The principles that apply to developing miRNA-based therapies remain the same as for other targeted therapies that take the path from drug target to drug. MiRNAs are likely to be high-potential drugs in the future.

MiRNAs influence a given biological system through the regulation of the target mRNAs. Therefore identification of miRNA targets is crucial for deciphering functional roles of the large numbers of miRNAs that are rapidly generated by the sequencing efforts. Currently, between 10 and 30% of genes are estimated to be regulated by miRNAs [22-24]. On average, miRNAs bind

to between 100 and 200 target sites on genes [25, 26], with some that have a few thousands interaction sites [27]. The number of known miRNAs has substantially increased over the last few years. The recent release 21 of the miRBase database [28-31] includes over 28 thousand miRNAs from 200+ species. Unfortunately, the annotation of their targets falls behind as only about one thousand miRNAs (4%) have validated targets. Moreover, the number of curated targets per miRNA is far lower than their estimated count. This motivates the development of high-throughput methods that predict miRNA targets. Dozens of computational miRNA target predictors have been developed since the first method was released in 2003 [32]. The underlying principle is to use annotated data generated by low-throughput experimental methods to build predictive models that can be used to perform high-throughput predictions for the miRNAs of interest that lack the experimental data. The current predictors differ on many aspects including their underlying predictive methodology, empirical evaluation, usability, popularity/impact, and predictive performance. Availability of many difficult-to-compare methods makes it challenging for the end user to select a proper tool and prompts the need for contributions that summarize and evaluate these methods to guide the users and to help the developers to revitalize this field.

The predictions generated by most of the current methods heavily depend on complementary Watson-Crick (WC) base pairing in the seed region, which encompasses the first eight nucleotides at the 5' end of miRNAs [23, 33]. They take the number of WC base pairs in the seed as the input (among some other inputs) and feed it into their pre-set scoring functions or learning models to generate the predictions. They rely on an assertion that the more WC base pairs are found in the seed, the more likely it is that the given miRNA interacts with the corresponding mRNA. Consequently, mRNA sites with more matches (WC base pairs) for a given miRNAs are predicted as targets more often than mRNA sites with fewer matches. Some of the current methods do not even predict targets with fewer than 7 WC base pairs in the seed; therefore, we define these neglected targets as non-canonical targets (<7 WC base pairs in the seed). Several biochemical studies provide evidence that miRNAs regulate non-canonical targets [34-36], and a few studies also reported that between 25% and 85% of targets are non-canonical, depending on a given type of the high-throughput experiments [37-39]. New computational approaches that improve predictive quality of non-canonical miRNA target prediction are therefore needed.

1.1 Thesis statements and goals

Motivated by the observations that miRNAs have gained a lot of attention during the past few decades, the objective of this thesis is to investigate the current miRNA target predictors and find their merits and disadvantages, and to build a new computational method that addresses these disadvantages. We formulated the following thesis statements:

1. The growth of the number of miRNAs is characterized by a fast pace. This is primarily because the high-throughput sequencing data are generated at a progressively lower cost and computational analysis of these data is relatively easy.
2. The current miRNA target predictors are very different in scope and usability (types of availability and ease of use).
3. The predictive quality varies across different computational methods for miRNA target prediction and can be improved.
4. Non-canonical miRNA targets are abundant.
5. Non-canonical miRNA targets can be accurately predicted using sequence-based methods, i.e., methods that require only miRNA and mRNA sequences as inputs.

We define three goals to address the aforementioned thesis statements:

1. **Analysis and discovery of miRNAs**. This goal addresses the thesis statement 1. We analyze the reasons of the fast growth of miRNA space by considering the cost and pace of generating and analyzing the miRNA sequencing data. We also propose, build and apply a computational platform for the analysis.

2. **Systematic review and comparative analysis of current computational miRNA target prediction methods.** This goal addresses thesis statements 2 and 3. We conduct systematic review of the current miRNA target predictors from both analytical and empirical perspectives to summarize this field and to analyze in-depth advantages and drawbacks of individual predictors. We provide insights for developers to design better prediction methods and for end users to select appropriate predictors. Our analytical description summarizes the scope, usability (availability and ease of use), popularity/impact, and predictive methodologies of the existing miRNA target predictors. Our empirical evaluation compares predictions at different levels of annotations for a representative set of current predictors.

3. **Development of an accurate and novel *ab initio* predictor of non-canonical miRNA targets.** This goal addresses thesis statements 4 and 5. We quantify the number of non-canonical miRNA targets and evaluate current methods for prediction of these targets. We design an accurate and novel predictor that takes only the sequences of miRNA and mRNA as its inputs with the goal to outperform the current predictors on the prediction of the non-canonical targets.

To summarize, our work provides insights for the end users to select an appropriate set of predictors for a given task at hand (a given miRNA) and for the developers to design and assess novel target predictors. Our new computational method is the first to accurately predict non-canonical miRNA targets from miRNA and mRNA sequences.

1.2 Thesis outline

In Chapter 2, we introduce the biological background concerning miRNA, miRNA-mRNA interaction and characteristics of the miRNA-target complexes, and computational background including predictive models and evaluation procedures. Since the amount of miRNAs grows so fast, in Chapter 3 we investigate the reasons for this growth; we also describe our platform for end-to-end processing of miRNA data generated by next generation sequencing and for prediction of new miRNAs. MiRNAs exert their functions through targeting mRNAs, so Chapter 4 reviews the current predictors, summarizes their properties, compares their predictive qualities and provides interesting relevant observations. Since we find that the current predictors are not suitable to accurately predict non-canonical miRNA targets, Chapter 5 describes our novel design that quickly and accurately predicts the non-canonical miRNA targets. Finally, the last chapter presents summary and conclusions, list of major contributions and findings, and an outline of possible future research directions.

Chapter 2

Background

2.1 MicroRNAs

MiRNAs are small (~22 nucleotides) endogenous noncoding RNAs. This section introduces the biogenesis of miRNA which includes two stages: production of miRNA sequences and targeting mRNAs by miRNAs. The production involves three steps including transcription, export, and post-transcriptional modifications, after which two more steps are used by the mature miRNAs to target mRNAs (see Figure 2-1):

Step 1. DNA (deoxyribonucleic acid) of miRNA is transcribed (copied) into RNA (ribonucleic acid), specifically into primary miRNAs (pri-miRNAs, about several hundred nucleotides long).

Step 2. The pri-miRNA is cut into 1-6 miRNA precursors (pre-miRNAs, about 70 nucleotides long) which are characterized by hairpin structures.

Step 3. The hairpin loop of the pre-miRNA is cleaved off and one pre-miRNA produces a pair of complementary (with imperfect WC pairing) miRNAs that are about 22 nucleotides long.

Step 4. The mature miRNA interacts with the RNA-induced silencing complex (RISC). The Argonaute protein in the RISC has two domains to hold the two ends of the mature miRNA to keep it straight [40].

Step 5. The interacting miRNA is used as a template to guide the complex to find the target position on the mRNA sequence that offers (to some extend) complementary WC base pairing. The argonaute then cleaves the mRNA at the binding position or inhibits its translation (a process that creates proteins) [41]. In either way, the corresponding protein cannot be synthesized.

The first two steps happen in the nucleus (organelle that contains most of the cell's genetic material), and then the product from nucleus – pre-miRNA is exported into cytoplasm (that contains all the other organelles in a given cell, except for the nucleus). The subsequent steps are localized in the cytoplasm. The first three steps involve the formation of miRNAs. At the end,

two miRNA strands are produced. Usually only one strand plays functional roles and the other one is degraded [42]. Sometimes both strands work as mature products, which means they both bind the target mRNAs. Our work primarily focuses on the last step.

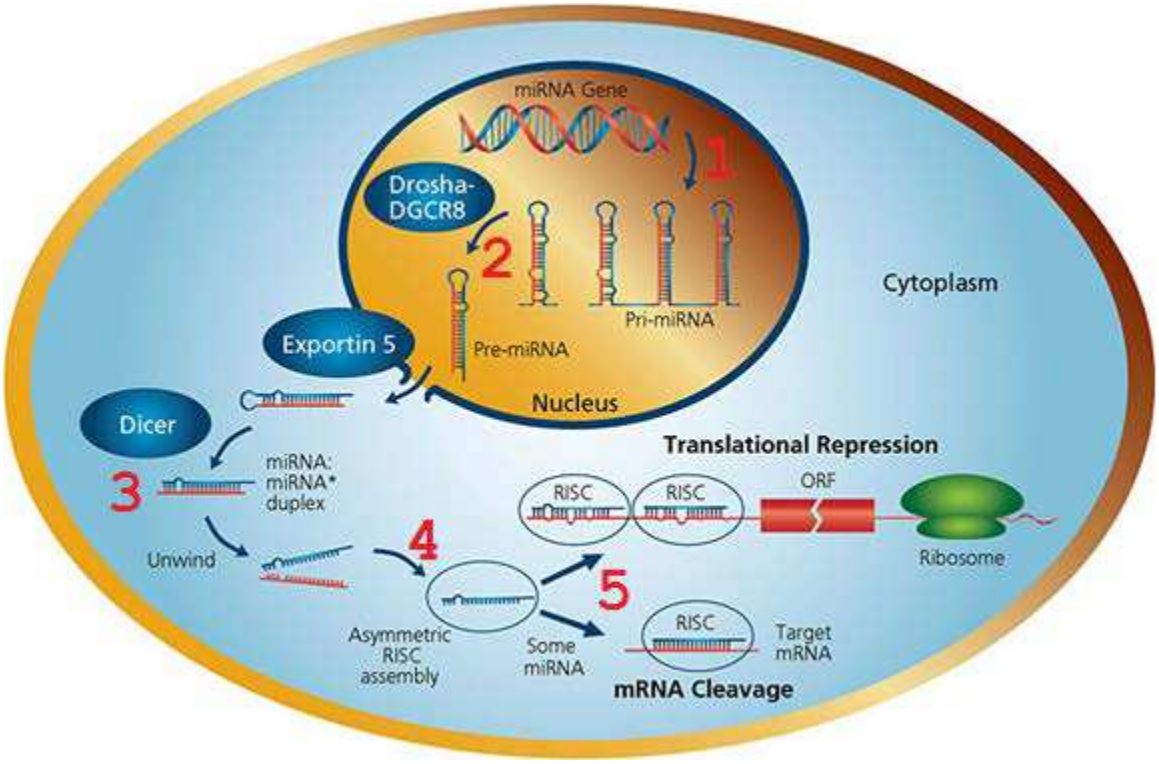


Figure 2-1 Biogenesis of miRNA and miRNA-mRNA interaction
Stage 1 that includes steps 1 to 3 describes the process of miRNA formation, and stage 2 that includes step 4 and 5 describes how miRNAs function, resulting in mRNA repression. Steps are shown as large red numbers.
The image was taken from miRNA pathway © SIGMA-ALDRICH (http://www.sigmaaldrich.com/life-science/functional-genomics-and-rnai/mirna/learning-center/mirna-introduction.html).

MiRNAs are now recognized as important regulators of a wide range of cellular processes. Understanding functions of miRNAs helps us to decipher the working of biological systems, such as cells or organisms. Moreover, miRNAs also act as an additional layer of gene regulation which can be dysregulated in diseases, i.e., miRNAs research has significant potential to study the pathogenesis of various disorders and diseases. To date, miRNAs have been used to study cardiovascular diseases [43], neurodegenerative diseases [44], metabolic diseases [45], and cancer [46], to name but a few. Finally, miRNAs have potential to be used in genetic therapeutics and provide assistance in the design of new drugs. For example, miRNA-based treatment of hepatitis C virus (HCV) infection has entered phase 2 clinical trials [47]. MiRNA

expression profiles are likely to become important diagnostic and prognostic tools in a near future [48], and miRNA replacement therapy is not far behind [18].

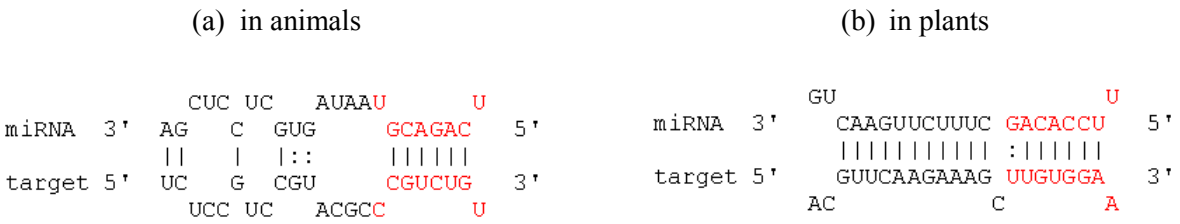


Figure 2-2 Illustration of miRNA-mRNA interaction/duplex
Two examples are given: in animals (panel (a)) and plants (panel (b)). Binding information (line 3) is drawn between miRNA and target mRNA sequences using '|' for WC base pair, ':' for GU wobble and ' ' for mismatch or gap. The seed region (first eight nucleotides) is in red. Binding nucleotides including both WC and GU are closer to the middle (lines 2 and 4) and nucleotides that are not bound (mismatch or gap) are further away from the middle (lines 1 and 5).

2.2 Characteristics of microRNA-mRNA interaction

This section further analyzes the step 5 that was discussed in section 2.1 which concerns the miRNA-mRNA interaction/duplex. Figure 2-2 illustrates the structure of a duplex between miRNA and mRNA sequences that is a consequence of the interaction. RNA is composed of four types of nucleotides: adenine (A), uracil (U), guanine (G) and cytosine (C). Nucleotides can pair with specific nucleotides through hydrogen bonds to form a more steady structure (to maintain lower free energy). The common base pairs are Watson-Crick (WC) base pairs (GC and AU) and wobble base pair (GU). The main biological characteristics of the miRNA-mRNA interaction which are often used in computational prediction include complementary base pairing, site accessibility, and evolutionary conservation**.** Other characteristics are not as commonly used and are usually utilized to improve specificity of predictions (reduce the number of false positive predictions).

2.2.1 Complementary base pairing

Complementary WC base pairing between miRNA and its target is the most common feature (characteristic/property) used to predict the potential miRNA targets. In contrast to the near-perfect base pairing in plants [49], animal miRNAs usually pair with their targets in a subset of

positions in the binding region [50]. However, base pairing in the seed region that is composed of the first eight nucleotides at the 5' end of miRNAs is particularly important. The major seed types include 6mer (six consecutive matches between 2^{nd} and 7^{th} positions from the 5' end of miRNA); 7mer-m8 (seven consecutive matches from 2^{nd} to 8^{th} position of miRNA); 7mer-A1 (7mer-A1 extends 6mer with an adenine (A) at the first position of target 3' end); and 8mer that combines 7mer-m8 and 7mer-A1.

2.2.2 Site accessibility

MiRNA target site accessibility is another common characteristic used in the target prediction. The accessibility is necessary since miRNA-mRNA interaction involves binding of a relatively large RISC [51]. This feature is usually quantified using content of adenine and uracil (AU content). Several studies found that enriched AU content in mRNA 3' untranslated regions (UTRs) is important for the interaction with miRNAs [52-54]. The site accessibility is also assessed using free energy that estimates stability of RNA sequences. Most predictors calculate the free energy of the miRNA-target duplexes. Some methods also calculate arguably more relevant relative energy which is the hybridization energy produced by miRNA-mRNA binding minus the disruption energy required by opening up the local mRNA structure of the target.

2.2.3 Evolutionary conservation

Evolutionary conservation refers to existence of similar or identical sequences of nucleic acids and proteins across different species that are often related taxonomically. Conservation of the miRNA targets is widely utilized to identify miRNA targets and to improve predictions by reducing the number of false positive predictions. Use of this feature is motivated by a premise which states that "similar" species should share common miRNAs and their targets. However, this leads to omission of non-conserved targets [55, 56]. The predictive value of the inclusion of the target conservation is an open question.

Some other characteristics are also used for predictive purposes but their effectiveness is controversial and requires further evidence to be more universally accepted. They include target abundance and binding patterns over the entire length of the miRNAs.

2.2.4 Target abundance

Presence of multiple target sites in a given mRNA is hypothesized to enhance the miRNA regulation [57, 58], and so some methods increase the propensity of binding to a target gene that they output when multiple sites are predicted on this gene.

2.2.5 Binding pattern

Some methods use specific patterns of binding in the miRNA-mRNA complexes that are found empirically in training data (dataset that is used to design the predictive model). These patterns consider certain positions outside of the seed regions that are empirically found to be important for the interaction. However, these patterns are relatively rarely employed since they vary depending on the training datasets used.

2.3 Data sources for the microRNAs and microRNA targets

MiRBase is the main microRNA database [59]. It started with 218 miRNA entries in 2002 and the number of entries has quickly increased to over 28 thousands miRNA sequences in 200+ species according to its latest release 21 in 2014. MiRNAs are found from deep sequencing data where small RNA sequences with high read counts are considered as potential miRNAs. Mature miRNA products are further verified through various biochemical experiments.

Table 2-1 Summary of databases of miRNA targets
Databases are sorted by their year of publication. "Duplex structure" indicates if the database provides the interaction information of the miRNA-target duplexes. "Validation methods" shows if the experimental methods that validated the miRNA targets are given. Unknown information is denoted by 'x'; 'NA' denotes that this information is not applicable since a given database did not focus on diseases.

Database	Year released	Duplex structure	# of miRNA-mRNA annotations	# of miRNAs	# of target genes	# of diseases	Validation methods	# of species		
								Animals	Plants	Viruses
miR2Disease	2008	No	3273	349	x	163	No	1	0	0
TarBase v6.0	2011	Yes	30597	706	14078	NA	Yes	6	1	1
miRecords v4	2013	Yes	2574	521	1637	NA	Yes	14	0	1
miRTarBase_4.5	2013	Yes	51460	1232	17520	NA	Yes	13	1	4
miRCancer	2014	No	2577	399	x	172	No	2	0	0

There are five popular databases of experimentally validated and curated miRNA targets, see Table 2-1. Only three of them provide information necessary to identify the miRNA-mRNA

duplexes: TarBase, miRecords, and miRTarBase. MiRTarBase version 4.5 stores the largest number of over 5000 miRNA-target duplexes [60] and also the largest number of non-functional miRNA-mRNA samples (mRNA genes that are validated not to interact with the given miRNAs). MiRecords includes only 2574 interactions [61]. TarBase's latest release v6.0 was substantially expanded compared to older versions; however, it does not provide the details of the interaction between miRNAs and mRNAs [62]. miR2Disease [63] and miRCancer [64] focus on selected diseases associated with miRNAs and also do not include information about the miRNA-mRNA duplexes.

2.4 **Prediction of microRNA targets**

Although there are several experimental methods to validate miRNA targets, they are relatively laborious and expensive. High-throughput predictions are needed to narrow down the list of potential miRNA targets and guide the biochemists to choose more likely targets for experimental validation. Computational prediction of miRNA targets have flourished over the last decade. Dozens of computational methods have been developed to predict the targets of miRNAs since the first predictor was proposed in 2003. Some predictors, especially the early ones, have been commonly used and the corresponding publications are cited over 1000 times [23, 24, 33]. These methods vary in the underlying predictive models and inputs, which is related to an observation that the training datasets were different and covered a diverse group of miRNAs binding mRNAs. Although the nature of miRNAs targeting mRNAs is still not entirely understood [41], some properties of these miRNA-mRNA interactions are known and used to provide useful information for the prediction. These characteristics were discussed in section 2.2.

The current miRNA target predictors also apply different predictive models. We divide these models into two categories: heuristic and empirical. Predictors are considered heuristic if they use manually developed (in an *ad-hoc* manner) screening or scoring models to filter targets or combine several input/features. This category is very popular because of the shortage of data for training and empirical design, which was especially true in the early stages of this field. Other advantages of heuristic models are convenience (ease) to set up, flexibility to integrate additional features, and algorithmic (runtime) efficiency. The second category uses an empirical approach to build and parameterize the predictive model using a training dataset. These models are designed by fitting predictions to known outcomes in the training dataset. They arguably can

better discover complex patterns that govern miRNA-mRNA interactions but they also rely on the quality of the training dataset, which is often plagued by problems with small size, and balance and distribution of the positive (functional/interacting) and negative (non-functional/not interacting) data. The functional miRNA-mRNA interactions are defined as those where the mRNA is down-regulated by the corresponding miRNA. Common types of predictive models in this field are logistic regression, support vector machine (SVM), genetic programming, Bayesian statistical modeling, and artificial neural networks. We introduce the most commonly used regression and SVM models that we utilize in our projects. These details are provided in section 2.5.1.

2.5 Background on computational methods

We utilize annotated data that is labeled with two classes to build and test prediction models. With the two classes, the annotations are labeled as positive (the interacting miRNA-mRNA pair) and negative (a pair that does not interact). To develop a predictive model the annotated data is divided into two datasets: training dataset and test dataset, which are independent (they do not include the same data). The objective is to train a model to maximize its predictive performance using the annotations on the training dataset. Once a desired level of predictive performance is reached on the training dataset, the same model is used to perform predictions on the test dataset and the predicted annotations are compared with the true annotation to assess the model and compare with other methods. The design of our predictive model is based on a "shotgun" approach where we generate a large number of features (numerical descriptors) from the input data (miRNA and mRNA sequences) which are potentially useful to separate positives from negatives. During the training process, we choose the prediction model type, its parameters, and a subset of features that provide highest predictive quality on the training dataset. This process is based on cross validation (details in Section 2.5.3), which simulates tests on the independent test dataset and aims to reduce chance of over-fitting the training dataset (i.e., generation of a model that fits too closely the training dataset). When testing on the test dataset, the predictive model converts the input data into the same set of features that were selected through training, feeds them into the model trained on the training dataset, outputs the predicted annotations (outcomes), compares the outcomes with the real labels, and evaluates the outcome using selected set of appropriate measures of predictive performance (details in Section 2.5.2).

2.5.1 Predictive models

As mentioned in Section 2.4, we introduce the most commonly used in this area of research regression and SVM models, which we consider in our design.

2.5.1.1 Logistic regression

The linear logistic regression [65], which is used to perform prediction of miRNA targets, often uses the least squared error criterion to parameterize the predictive model (compute values of coefficients). Given the outcome $\mathbf{y} \in R^{t\times 1}$, which is the annotation of miRNA-mRNA pairs (binding or functional vs. non-binding or non-functional) and a set of input features $X \in R^{t\times n}$ (computed from miRNA and mRNA sequences), where *t* is the number of miRNA-mRNA pairs, *n* is the number of features used in the regression model, the criterion to solve the regression model is defined as:

$$\min_{\mathbf{r}} \left(\left\| \mathrm{X}\mathbf{r} - \mathbf{y} \right\|_2^2 + \beta \left\| \mathbf{r} \right\|_2^2 \right) \tag{2-1}$$

where $\mathbf{r} \in R^{n\times 1}$ are coefficients and β is a regularizer used to adjust the trade-off between the squared error (the first part) and regularization (the second part). Introduction of β helps to avoid over-fitting into the training dataset. To control the increased error brought by the use of the regularizer, β is usually set to a small value.

2.5.1.2 Support vector machine

SVM was developed by Vapnik [66] and gained popularity in recent years. SVM solves a convex optimization problem, which means that the model is obtained by finding a unique global optimum. This is an advantage over some other techniques, such as neural networks, that may get stuck in local minima/maxima. This model does not require a large number of training samples compared to the number of features (inputs) to secure good predictive performance. Given the labels/annotations $\mathbf{y} \in R^{t\times 1}$ and inputs/features $X \in R^{t\times n}$, the SVM model is a classification hyper-plane in the mapped higher dimensional space (generated with help of a kernel function) with a margin of maximal width that is generated by solving a convex quadratic programming problem [67]:

$$\min\left\{J(\mathbf{w},b)=\frac{1}{2}\|\mathbf{w}\|^2\right\},\ \text{s.t.}\ \ y_i\left(\langle\mathbf{w},\mathbf{x}_i\rangle+b\right)\geq 1 \tag{2-2}$$

where, $\mathbf{w}$ is the normal to the hyper-plane and $|b|/\|\mathbf{w}\|$ is the perpendicular distance from the hyper-plane to the origin, and $\|\cdot\|$ is the Euclidean norm. The above formulation cannot be solved when the input data cannot be separated by a hyper-plane, and thus a set of positive slack variables in the constraints is introduced:

$$\min\left\{J(\vec{w},\xi_i,b)=\frac{1}{2}\|\vec{w}\|^2+C\sum_{i=1}^{N}\xi_i\right\}\quad \text{s.t.}\ \begin{cases} y_i\left(\langle\vec{w},\vec{x}_i\rangle+b\right)\geq 1-\xi_i \\ \xi_i\geq 0,\ \ \forall i=1,2,\cdots,N \end{cases} \tag{2-3}$$

where, $\sum_{i=1}^{N}\xi_i$ is an upper bound on the number of training errors and C is a penalty parameter. A larger C corresponds to a higher penalty to errors and a stricter constraint, which results in a finer-described hyper-plane, a smaller margin width, and potentially higher classification accuracy on the training dataset; but it also results in a potentially reduced generalization into the unseen samples (outside of the training dataset). The above definition can be converted into Lagrangian dual problem:

$$\begin{aligned} &\max\left\{L(\alpha_i,\alpha_j)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\langle\vec{x}_i,\vec{x}_j\rangle\right\} \\ &\text{s.t.}\begin{cases} \sum_{i=1}^{N}\alpha_i y_i=0 \\ \alpha_i,\ \alpha_j\in[0,\ C],\forall i,j=1,2,\cdots,N \end{cases} \end{aligned} \tag{2-4}$$

where α_i and α_j are Lagrange Multipliers. If $\alpha_i\neq 0$, the corresponding $\mathbf{x}_i$ is called support vector (SV). The set of SVs determines the classification hyper-plane [68]. SVM is often applied with kernel to extend linear discriminant machines into the nonlinear domain through dot products in a higher dimensional feature space. The SVM with a kernel is expressed as

$$\begin{aligned} &\max\left\{L(\alpha_i,\alpha_j)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)\right\} \\ &\text{s.t.}\begin{cases} \sum_{i=1}^{N}\alpha_i y_i=0 \\ \alpha_i,\alpha_j\in[0,C],\forall i,j=1,2,\cdots,N \end{cases} \end{aligned} \tag{2-5}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \right\rangle$$

where the operator $\langle \cdot \rangle$ means inner product, $\Phi(\cdot)$ maps the data from a low dimension to a high dimension, and $K(\cdot)$ is the kernel function. The form of kernel should satisfy the Mercer's condition [69]. Gaussian Radial Basis Function (RBF) kernel, which is one of most widely used kernels, is given as follows:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\| \mathbf{x}_i - \mathbf{x}_j \|^2}{2\sigma^2} \right), \ \sigma \in \mathbb{R}^+ \tag{2-6}$$

2.5.2 Measures of predictive quality

Predictions of miRNA-target are assessed at the duplex (to predict whether a given fragment on mRNA interacts with a given miRNA) and the gene (to predict whether a given mRNAs interacts with a given miRNA) levels. The predictions usually take two forms: 1) a ***binary value*** that indicates whether a given miRNA-target pair (at either the duplex or the gene level) is predicted bound (functional) or not; and 2) a ***real value*** that quantifies propensity of the corresponding binding. The binary predictions are generated by thresholding the real valued outcomes, i.e., pairs with scores above a given threshold are assumed function and below as non-functional. The ***binary predictions*** are assessed by one or more of the following seven measures:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2-7}$$

$$Specifity = \frac{TN}{TN + FP} \tag{2-8}$$

$$Precision = \frac{TP}{TP + FP} \tag{2-9}$$

$$SNR+ = \frac{TP}{FP} \tag{2-10}$$

$$SNR- = \frac{TN}{FN} \tag{2-11}$$

$$PNR = \frac{TP + FP}{TP + FN} = \frac{Sensitivity}{Precision} \tag{2-12}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2\text{-}13)$$

where *TP* (true positives) and *TN* (true negatives) are the counts of correctly predicted functional and non-functional genes/sites respectively, and *FP* (false positives) and *FN* (false negatives) are the counts of incorrectly predicted functional and non-functional genes/sites respectively. Signal-to-Noise Ratio of correctly over incorrectly predicted functional targets (*SNR*+) was calculated in previous works [32, 33, 70-72]. We also introduce the *SNR* of correctly over incorrectly predicted non-functional samples (*SNR*-) to complement the *SNR*+ criteria. Further, since the counts of true functional and non-functional samples are skewed (highly different), we normalize the *SNR* value as follows:

$$SNR\_duplex+ = \frac{TP\_duplex}{FP\_duplex} \Big/ \frac{P\_duplex}{N\_duplex} \quad (2\text{-}14)$$

$$SNR\_duplex- = \frac{TN\_duplex}{FN\_duplex} \Big/ \frac{N\_duplex}{P\_duplex} \quad (2\text{-}15)$$

$$SNR\_gene+ = \frac{TP\_gene}{FP\_gene} \Big/ \frac{P\_gene}{N\_gene} \quad (2\text{-}16)$$

$$SNR\_gene- = \frac{TN\_gene}{FN\_gene} \Big/ \frac{N\_gene}{P\_gene} \quad (2\text{-}17)$$

where *P_duplex* and *N_duplex* are the numbers of true functional and non-functional duplexes; and *P_gene* and *N_gene* are the counts of true functional and non-functional genes respectively. We also assess the total count of predicted functional targets using Predicted-to-Native positive Ratio (*PNR*) = predicted_functional_count/true_functional_count. *PNR* shows whether a given predictor over-predicts (*PNR* value is larger than 1) or under-predicts (*PNR* value is less than 1) the number of functional miRNA targets. The Matthews Correlation Coefficient (*MCC*) provides a balanced measurement of the predictive quality of the functional and non-functional predictions; this measure is appropriate for the assessment of skewed dataset. The values of *MCC* range between -1 to 1 with 0 denoting random prediction and higher values denoting more accurate predictions.

The ***real valued*** predictions are computed at the gene level as the sum of probabilities of all predicted targets in that gene. The Receiver Operating Characteristic (*ROC*) curve, which

represents relation between true positive rate (*TPR*) = *TP*/(*TP*+*FN*) and false positive rate (*FPR*) = *FP*/(*FP*+*TN*), is used to evaluate the predicted probability. The *TPR* and *FPR* values are computed by thresholding the real valued predictions using every unique predicted value as the threshold. The *ROC* curves reflect a trade-off between sensitivity and specificity, providing comprehensive information about the performance of the model. The predictive quality is assessed with the Area Under *ROC* Curve (*AUC*) that ranges between 0 (for a prediction model that does not correctly predict any of the positives) to 1 (for a perfect prediction model), with 0.5 denoting a random predictor [73]. Except for the *PNR* and *SNR* of the non-functional predictions that we introduce and the normalization of the *SNR* values, the other criteria were used before [74-79].

2.5.3 Cross validation

We use cross validation on the training dataset to design prediction models (e.g., to select a subset of features, choose the best-performing type of predictive model, optimize parameters of the selected model). This methodology helps to prevent over-fitting the training dataset, and is often used to assure that the estimates of predictive quality on the training dataset (based on cross validation) transfer into the independent test dataset [80]. In cross validation, we partition at random a given training dataset into *k* equally-sized subsets/folds (hence name *k*-fold cross validation). Then, one of the subsets is used to test a model that is trained using the remaining *k*-1 subsets, and this is repeated *k* times so that each subset is used once to perform the test. We combine the results from the *k* tested folds together, which corresponds to performing tests on all training instances.

2.5.4 Statistical tests

We assess statistical significance of differences in predictive performance for a given pair of prediction models. A statistical test is a method with a pre-defined null hypothesis H_0 that assumes that two sets of data points come from the same population against an alternative hypothesis H_1. A probability *p*-value of that the null hypothesis is actually true is used to accept or reject hypothesis H_0. If *p*-value is smaller than a certain significance level, the hypothesis H_0 is rejected. In this thesis, statistical tests are based on 10 repetitions of randomly chosen 50% of data from two compared datasets. The *p*-value level is set at 0.05. There are different tests for

different types of data. We use Student's t-test [81] if the distributions of the data points are normal, otherwise we utilize the Mann-Whitney *U* test [82]. Distribution type is verified using the Anderson-Darling test [83] with the *p*-value of 0.05.

2.5.4.1 Student's t-test

This test assumes that values of data points follow normal distribution. The null hypothesis assumes that the means of the two compared groups are equal. The test is defined by the following equation:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2-18}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_{X_1}^2}{n_1} + \frac{S_{X_2}^2}{n_2}}} \tag{2-19}$$

where:

$S_{X_1X_2} = \sqrt{\frac{(n_1-1)S_{X_1}^2 + (n_2-1)S_{X_2}^2}{n_1+n_2-2}}$ is an estimator of the standard deviation of the two sets of data points;

$S_{X_i}^2 = \frac{1}{n_i}\sum_1^{n_i}(x_k - \bar{X}_i)$ is the variance of set i $(i = 1, 2)$;

$\bar{X}_i = \frac{1}{n_i}\sum_1^{n_i} x_k$ is the mean of set i $(i = 1, 2)$;

x_k $(k = 1, 2, \dots, n_i)$ is the k^{th} value of set i $(i = 1, 2)$;

n_i is the number of data points in set i, $(i = 1, 2)$.

Equation 2-18 is for two sets with the equal variance and equation 2-19 is for unequal variances. Once the *t*-value is calculated, the significance can be found using a table of *t*-values from the Student's *t*-distribution. The table is composed of critical values defined by significance level (*p*-value) and degree of freedom (DF). If the calculated *t*-value is larger than or equal to the critical value given in the table, then the compared two sets of data points are significantly different, i.e., the null hypothesis H_0 is rejected at the level of significance *p*-value; otherwise they are not significant different and H_0 is accepted.

2.5.4.2 Mann-Whitney *U* test

This is a nonparametric test and makes no assumptions about the distributions of the values of sets of points being assessed. Thus, this test is preferred over the student *t*-test for data with non-normal distributions. The null hypothesis assumes that the medians of the two compared sets are equal. The test is defined as:

$$U = \min_i \left\{ R_i - \frac{n_i(n_i+1)}{2} \right\} \tag{2-20}$$

where:

R_i is the sum of the ranks in one set of data points $i, (i = 1, 2)$;

n_i is the number of data points in set $i, (i = 1, 2)$.

Once the *U*-value is calculated, the significance can be found using a Mann-Whitney table. The critical values in the table are defined by significance level (*p*-value) and the number of data points in the two groups. If the calculated *U*-value is larger than or equal to the critical value given in the table, then the compared two sets of data points are significantly different, i.e., the null hypothesis H_0 is rejected at the level of significance *p*-value; otherwise they are not significant different and H_0 is accepted.

2.5.4.3 Anderson-Darling normality test

This test is used to test whether values of a set of data points come from a given probability distribution. As we focus on normal distribution, the null hypothesis assumes the given set of values is normal. The test is defined by the following equation:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\ln \Phi(Y_i) + \ln\left(1 - \Phi(Y_{n+1-i})\right)\right), Y_i = \frac{X_i - \bar{X}}{\sigma} \tag{2-21}$$

where:

X_i is the i^{th} value of the given set of data points *X*;

$\bar{X}$ is the mean of *X*;

σ is the standard deviation of *X*;

$\Phi(Y_i)$ is a cumulative distribution function of Y_i for normal distribution;

n is the number of data points in X.

In our case, both the mean $\bar{X}$ and variance σ are unknown. Once the A^2 is calculated, the significance can be found using an Anderson-Darling table. The critical values in the table are defined by significance level (p-value) and the number of data points. If the calculated A^2 value is larger than or equal to the critical value given in the table, then the set of data points is not normal, i.e., the null hypothesis H_0 is rejected at the level of significance p-value; otherwise the values of the data points are assumed to be drawn from normal distribution and H_0 is accepted.

Chapter 3

Analysis and Discovery of MicroRNAs

3.1 Motivation

Sequencing is a technique to determine the primary structure (sequence) of DNAs, RNAs or proteins. Since the first method was proposed in 1970 [84], much progress has been made to make this technique faster and more affordable. The advent of the first high-throughput sequencing method in 2005 [85] brought a substantially increase in the number of sequencing studies. Several companies developed their own methods, designed corresponding machines, and put them on the market. This high-throughput sequencing is often called next generation sequencing (NGS). It can replicate and screen a whole genome, transcriptome, or a particular type of genes in a relatively fast manner (hours to days), and measure their expression levels. NGS has a great potential in biological, medical and clinical research, and it has been widely used to study genes, gene evolution and regulation. For miRNA sequencing, it is often used to profile miRNA expression levels and find novel miRNAs. NGS generates massive amounts of data that have to be computationally and automatically cleaned, processed, annotated and analyzed. This requires availability of specialized and integrative software pipelines to translate the sequencing data into results that are generated automatically in an easy-to-comprehend manner. We found the counts of miRNAs in miRBase have enjoyed a quadratic growth in the last decade (Figure 3-1) and we hypothesize that this growth is due to the low cost of the NGS that continually decreases over time (Figure 3-2), easiness to build computational platforms, and their wide availability. We introduce several existing platforms in Section 3.2 and describe our own platform in Section 3.3. We also apply our platform in a study related to endoplasmic reticulum (ER) stress in Section 3.4.

The work in Section 3.3 and 3.4 has been published in refs [86, 87] with collaboration with Dr. Michalak's group. The development of our computational pipeline allowed his group to formulate hypothesis related to ER stress, which concerns finding of a specific miRNA from a large pool of short RNA reads that targets specific genes of interest. We found this miRNA and

also the corresponding targets. The subsequent experimental studies were done by the group of Dr. Michalak, and thus we do not discuss these details in this thesis.

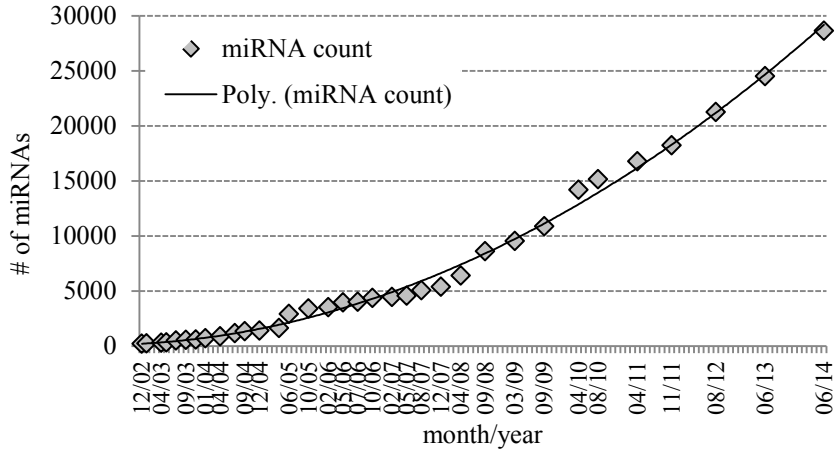


Figure 3-1 Growth of the number of miRNAs in the last decade
Diamonds give the number of miRNAs in miRBase for a date shown in the *x*-axis. Black line plots the second-order polynomial fit.

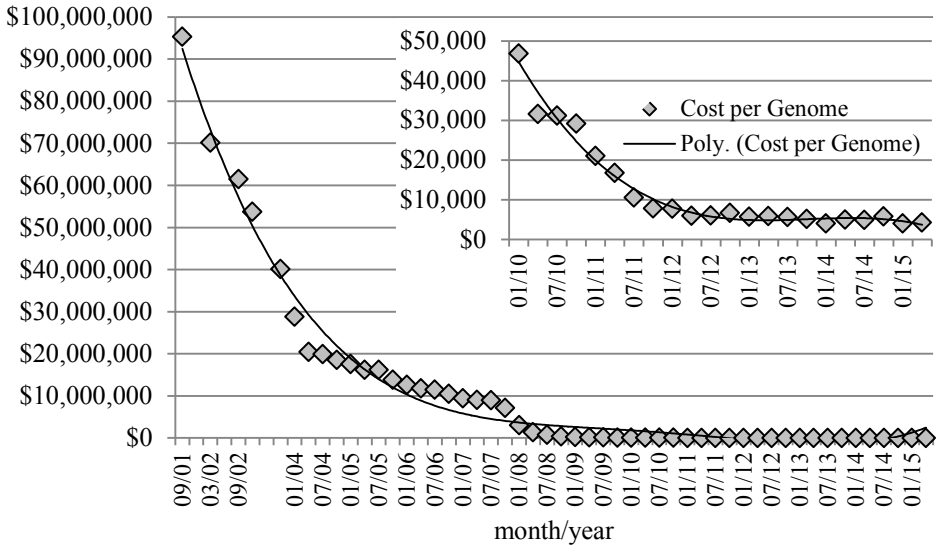


Figure 3-2 Cost of generating sequencing data per genome
The diamonds give the cost of generating sequencing data per human genome for a date shown in the x-axis. Black lines plot the fourth-order and third-order polynomial fit in the main figure and the insert. The insert in the up-right corner zooms in on the cost in the past five years. The data were taken from National Human Genome Research Institute (http://www.genome.gov/sequencingcosts/)

3.2 Current platforms for next generation sequencing data

The large amount of miRNA data generated by NGS cannot be processed manually. Since this sequencing technique was applied to other RNAs much earlier than to miRNAs, there have been

existing computational tools out there. They are not designed specifically for miRNAs but they provide foundation for developers to easily update and assemble the tools necessary for analysis of miRNA data. By now, at least a dozen software integrated pipelines have been developed to clean, process, annotate and analyze miRNA sequencing data, see Table 3-1.

Table 3-1 Analysis and comparison of existing miRNA processing pipelines and described here new pipeline
"x" denotes availability of a specific functionality.

Pipelines	Year	Webserver	Standalone package	Detection of known miRNAs	Prediction of novel miRNAs	Differential expression	Prediction of miRNA targets
CAP-miRSeq [88]	2014			x	x	x	
Kraken [89]	2013		x	x		x	
CPSS [90]	2012	x		x	x	x	x
miREvo [91]	2012		x	x	x	x	
Ours	2011			x	x	x	x
DARIO [92]	2011	x		x	x		
wapRNA [93]	2011	x	x	x	x	x	x
deepBase [94]	2010				x		
MIReNA [95]	2010		x	x			
miRNAkey [4]	2010		x	x		x	
DSAP [3]	2010	x		x		x	
SeqBuster [96]	2010	x	x	x		x	x
MAGIA [97]	2010	x		x			x
mirTools [98]	2010	x		x	x	x	
MiRExpress [99]	2009		x	x			
miRanalyzer [100]	2009	x	x	x	x	x	
miRDeep [101]	2008		x	x	x		

These pipelines provide different services; the baseline is to detect known miRNAs which can be done by each pipeline except for deepBase [94]. Nine and ten out of sixteen pipelines can also predict novel miRNAs and calculate differential expression levels between samples. CAP-miRSeq [88], CPSS [90], miREvo [91], wapRNA [93], mirTools [98] and miRanalyzer [100] perform both functions. Eight of these platforms are available as webservers, which is convenient for typical users (biologists and biochemists) who have little knowledge of computing. Nine platforms are available as standalone packages, which benefits advanced users who can assemble customized computational platform that satisfy requirements of their projects. WapRNA [93], SeqBuster [96] and miRanalyzer [100] are available both as standalone and webservers. Moreover, CPSS [90], wapRNA [93], SeqBuster [96] and MAGIA [97] also integrate target predictions. Overall, the most comprehensive to date platforms for processing the miRNA data are CPSS [90] and wapRNA [93]. They provide detection of known and

novel/putative miRNAs, perform statistical analysis of differential expression, and predict targets. We acknowledge that third-party tools can be used to implement the above functions, such as detection of known miRNAs (comparing short RNA sequences to contents of miRNA databases) and differential analysis. These tools would be used to post-process the results generated by some of the pipelines that are missing such analysis. However, this would be less convenient for the end-user and thus here we consider the pipelines that can fully automatically process the raw sequencing data to produce a complete set of results.

3.3 Our pipeline for processing microRNA data from NGS

We collaborated with Dr. Michalak's group on a project that involved processing of miRNA sequencing data in 2011. At that time CPSS was not published yet and wapRNA did not allow users to change parameters, e.g., degree of match to genomes and minimum reads count. We note that we completed this project in 2011, which is before the last four methodologies from Table 3-1 were introduced, and since then we did not update our pipeline. Our project required detection of known miRNAs, prediction of novel miRNAs, ability to compute differential expression, and prediction of miRNA targets. Since we could not find a suitable fit to our requirements, we build a new platform that integrated multiple computational tools. The flowchart of the pipeline is shown in Figure 3-3. A pool of short RNAs (reads) are cleaned up in three steps using the FastQC program (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc): 1) the sequencer's adapters are trimmed at the 3' end using the Btrim program [102]; 2) the continuous ploy-A/C/G/T/N at the 5' end are removed; 3) unique sequences between 18 and 29 nucleotides in length are retained together with their counts. Next, the remaining short reads go through four filtering steps, where sequence alignment is performed using the Bowtie program [103] assuming perfect match: 1) sequences that are not matched to the expected genome (the species that the miRNA data come from) are assumed to be contaminations and are discarded; 2) sequences that are matched to repetitive DNAs from Repbase [104] (uploaded from http://www.girinst.org/server/RepBase/) and non-coding RNAs, including transfer RNAs, ribosomal RNAs, small nuclear RNAs, and small nucleolar RNAs, from Rfam [105] (using build 10.0 from http://rfam.janelia.org/) are labeled as non-miRNAs and are removed; 3) short RNAs which are identical to the known miRNAs in miRBase [30, 31] are assigned as miRNAs and set aside; 4) the remaining short RNAs are processed to find putative miRNAs. Two putative

miRNA precursor sequences of the remaining short reads (one with 10 nucleotides upstream and 70 nucleotides downstream, assuming that miRNA is at the 5' arm of the RNA hairpin, and the other with 70 nucleotides upstream and 10 nucleotides downstream, assuming that miRNA is at the 3' arm) are processed by the MIREAP program to select those that have hairpin structure. The hairpin-like reads are folded using RNAfold [106] to select those with a minimum free energy below -25 kcal/mol. Finally, the short RNAs whose precursors satisfy the above requirements are clustered by the common precursor. MiRNAs on both arms of the hairpin are considered. Each cluster represents one putative miRNA, and its sequence is set to be the most frequent or abundant sequence in a given cluster. The abundance for each putative miRNA is calculated as a sum of abundance of all (similar) reads in this cluster. The purpose of our pipeline was not to serve the public but rather to facilitate a particular project that is discussed in section 3.4. Therefore, we limit the evaluation only to the scope of this project. We do not follow up to upgrade our pipeline because it was superseded by more recent pipelines, which by now are likely to be more suitable for current projects.

Thanks to availability of relevant computational tools, such as FastQC, Bowtie, MIREAP, and RNAfold, we were able to fairly easily build a new pipeline which fits requirements of our project. This experience suggests that it is easy to analyze miRNA sequencing data, even if none of the existing platforms can be used. This observation and the availability of many computational platforms contribute to the rapid growth of miRNA counts.

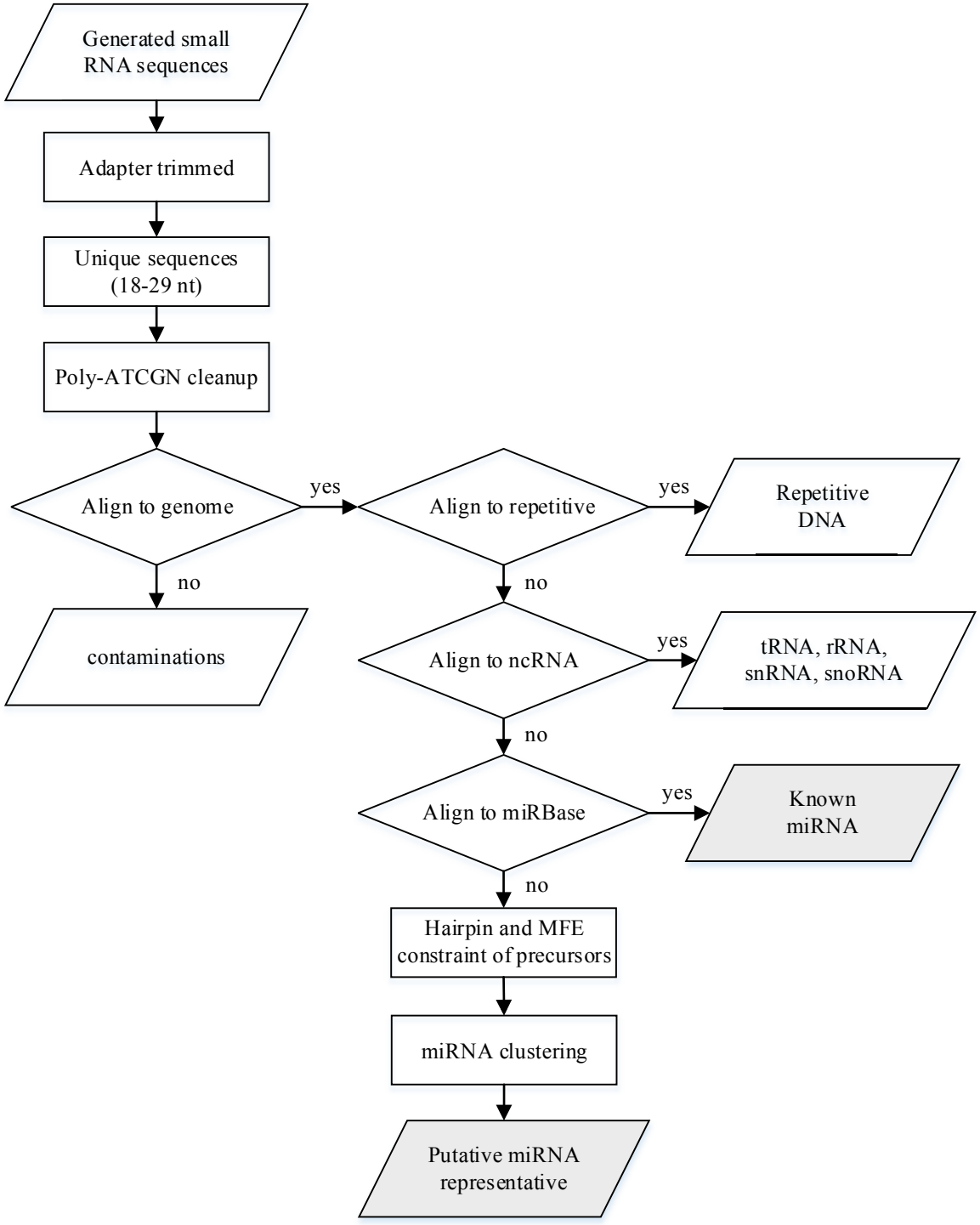


Figure 3-3 Flowchart of our pipeline for processing miRNA sequencing data
Data is shown in parallelograms. Data processing is shown with rectangles. Decisions are given with diamonds. The final outputs are shaded.

3.4 Application of our pipeline to find significant microRNAs in ER stress

The disruption of the energy or nutrient balance triggers endoplasmic reticulum (ER) stress. This application focuses on discovery of novel mechanisms induced by disrupted ER calcium homeostasis using miRNA deep sequencing.

3.4.1 Deep sequencing analysis

Wild-type NIH-3T3 mouse fibroblasts cells and NIH-3T3 cells treated with 500 nM thapsigargin (that raises cytosolic/intracellular calcium concentration) were cultured. Total RNA was harvested and used for deep sequencing on the Illumina GAIIx sequencer. We analyzed the small RNA reads generated from the sequencer using our pipeline from section 3.3. We used National Center for Biotechnology Information (NCBI) mouse genome (using build 37.2 from http://www.ncbi.nlm.nih.gov/projects/genome/guide/mouse/) to remove contaminations. The known miRNAs were tagged using release 17 of miRBase. Finally, all known miRNAs with counts below 5 were removed.

3.4.2 Application of the finding from sequencing in ER stress

Bioconductor package edgeR [107] was applied to determine whether a given miRNA was differentially expressed between the wild-type and thapsigargin treated groups. The miRNAs were sorted by the adjusted *p*-values, which were computed using trimmed mean of M values (TMM) normalization and tagwise dispersion. All miRNAs with adjusted *p*-value<0.5 were annotated with their (putative) target genes and considered for experimental validation. The experimentally validated targets were collected using miRecords database [61]. Because the number of experimental annotations was relatively low, we turned to computational prediction methods for help. We observed that current predictors provide quite diverse predictions (i.e., they predict a substantially different numbers of targets that are characterized by a relatively small overlap. Moreover, we could not find a source that would identify well or best performing methods. Thus, we used three popular target predictors [108]: TargetScan [23], DIANA-microT [109], and RepTar [110] and made an assumption that targets that are predicted by multiple methods are more reliable. These difficulties with establishing a protocol to predict miRNA targets motivated us to investigate miRNA targets predictors, which we describe in the next Chapter.

Among all analyzed miRNAs, abundance of mmu-miR-322-5p was found to be reduced by over 43% by depletion of calcium stores, and was predicted to target Pdia6 by both TargetScan and DIANA-microT. Dr. Michalak's group has experimentally validated the fact that the reduced abundance of miR-322-5p increases the stability of Pdia6 in both mice and worms. Their experimental validation has also shown that Pdia6 interacts and enhances IER1α (inositol-requiring enzyme 1α) activity. Activation of IER1α leads to the generation of transcription factor XBP1 (X-box binding protein 1), which controls the quality and folding of proteins. Together, ER calcium, Pdia6, IRE1α and miR-322-5p function in a dynamic feedback loop, which is important in a pathway to reestablish homeostasis of the ER. The work described in this section was done by Zhenling Peng and Jody Groenendyk [86, 87].

3.5 **Conclusions**

We analyzed data across historical releases of the miRBase database and found that the number of included miRNAs is characterized by a quadratic growth in the last decade. This number has grown substantially especially after 2005, which is the year when the next generation sequencing was deployed. We hypothesize that this growth is related to the development of high-throughput sequencing technology and availability of computational tools that can be used to analyze the resulting data. Our experience demonstrates that developers of such computational platforms for end-to-end processing of miRNA data can borrow mature computational tools for analysis of RNAs and relatively easily assemble them into a pipeline that satisfies requirements of their projects. We implemented such pipeline for a collaborative project. Our pipeline finds known and novel miRNAs, measures their expression levels, and evaluates differential expression among samples. We applied this pipeline to a project in ER stress, which resulted in finding a new mechanism for reestablishing homeostasis of the ER based on regulation via a specific miRNAs. This success shows that our easy-to-put-together platform generates useful results. To summarize, we conclude that emergence of the low-cost high-throughput sequencing, easiness to develop relevant computational tools, and availability of numerous ready-to-use tools fueled the rapid growth of the number of miRNAs. We note that although our pipeline was designed and applied to miRNA sequencing data in mouse, it can be easily extended to other species by replacing a few databases that it utilizes.

Chapter 4

Systematic Review and Comparative Analysis of the Current MicroRNA Target Prediction Methods

In the previous chapter we found that the growth of the number of miRNAs is high and likely will continue to be high given the availability of NGS and computational tools that process the corresponding data. Once these miRNAs are found/discovered, their function needs to be deciphered based on finding their targets. Although experimental methods can be used to validate miRNA targets, they are relatively very slow (low-throughput) and there are many potential targets to consider. As we show in Section 2.3, there are already over 28 thousands miRNAs in the miRBase database. However, only 1232 miRNAs have experimentally validated targets in the largest miRNA target database – miRTarBase, which accounts for 4.3% of all known miRNAs. This motivates the need for fast and high-throughput computational methods that can be used to find targets for a large number of miRNAs. When we integrated current target predictors into our pipeline, we found that they required different formats of inputs, provided different numbers of targets, and their predicted targets virtually did not overlap. This prompted us to take a closer look at the various methods that predict miRNA targets. Therefore, this chapter reviews the current computational miRNA target prediction methods. This work was published in ref. [111].

4.1 Motivation

Dozens of computational miRNA target predictors, which find targets from the mRNA and miRNA sequences, have been developed since the first method was released in 2003 [32]. The underlying principle is to use data generated by (usually low-throughput) experimental methods to build predictive models, which in turn can be used to perform high-throughput predictions for specific miRNAs of interest that lack the experimental data. The results generated by these (base) predictors can be filtered or combined together by meta predictors, i.e., methods that refine predictions of the base methods such as Pio's approach and myMIR [112, 113]. However,

the meta predictors often lack integration with the base predictive models (they were developed separately from the base methods and require manual collection of the predictions from the base methods) and they rely on availability of results generated by multiple base methods which makes them more challenging to use. The targets can be also predicted computationally by ranking the gene expression or CLIP-based data but in this case the inputs are the experimental data, which limits their applications. We focus on the computational miRNA target predictors that require only the knowledge of the miRNA and mRNA sequences (sequence-based miRNA target prediction), excluding the meta methods.

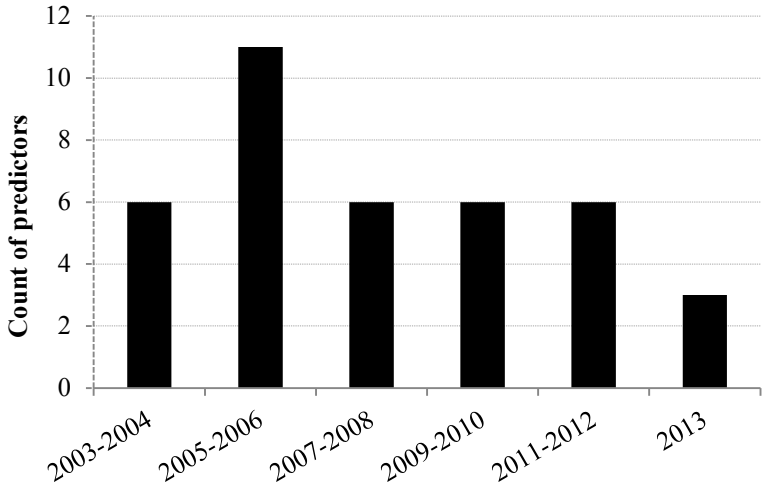


Figure 4-1 Count of miRNA target predictors published since 2003
Data is presented biannually.

The field of sequence-based miRNA target prediction has reached maturity, as evidenced by the declining trend in the development efforts (Figure 4-1). After the initial spike in 2005 when 8 methods were developed, more recent years have seen on average only three new methods per year. These predictors differ on many aspects including their underlying predictive methodology (type of predictive model they utilize; mechanistic details of miRNA-mRNA interaction that they consider including use of complementarity of base pairing, site accessibility, and evolutionary conservation), empirical evaluation (datasets and evaluation procedures), usability (availability and ease of use), popularity and impact, and predictive performance. Availability of many difficult-to-compare methods makes it challenging for the end users to select a proper tool and prompts the need for contributions that summarize and evaluate these methods to guide the users and to help the developers to revitalize this field.
Table 4-1 compares existing reviews of the miRNA target predictors based on the inclusion of discussion and analysis of the abovementioned aspects. We observe that these reviews summarized the latest miRNA target predictors at the time of their publication and compared or at least described the methodology utilized by these predictors. Most of these contributions also

discussed availability of predictors and some aspects of their usability, focusing on the species that they were designed for. However, other important aspects of usability, such as the number of input parameters (that determines flexibility of use for an expert user), the format of the input miRNAs and mRNA genes, the ability to predict for novel miRNA sequences, the format of the outputs, and the number of predicted targets (which differs substantially between methods) were omitted. They also neglected to discuss popularity and impact of the predictors and details concerning their evaluation. Only three relatively older reviews provided comparative evaluation. The first review by Rajewsky assessed nine methods on 113 experimentally annotated miRNA-target pairs, but only in *Drosophila* [114]. Review from 2006 by Sethupathy [115] used a small set of 84 annotated miRNA-target pairs and lacked assessment on the non-functional pairs (whether these methods can correctly recognize lack of interaction). The latest comparative review from 2009 by Alexiou [116] utilized 150 miRNA-target duplexes but considered only relatively old methods that were published in 2007 or earlier. Moreover, the evaluation criteria included only sensitivity and precision, which does not cover quality of prediction of the non-functional pairs. To summarize, prior reviews of the sequence-based miRNA target prediction methods suffer from lack or limited and outdated empirical evaluation, inclusion of a relatively small set of predictors, lack or shallow treatment of certain aspects, such as usability and impact of the prediction methods, evaluation procedures, and practical insights for the end users and developers.

Table 4-1 Summary and comparison of reviews of miRNA target predictors

The reviews are sorted by their year of publication. 'x' denotes available and 'o' means partially available functionality. We considered analytical and empirical components. Analysis of the analytical component includes the number of covered methods, the year when the latest included predictor was proposed, and whether the methodology, evaluation, usability, and impact dimensions were discussed. 'Methodology' concerns discussion of the inputs/features utilized by the prediction models and types of these models. 'Evaluation procedure' concerns review of the datasets and evaluation procedures for designing and assessing the predictors. Assessed usability is defined in terms of analysis of 'availability' and 'ease of use'. The former provides information about the availability of the methods to the end users, usually in the form of standalone software and/or a webserver. The ease of use focuses on other aspects of usability including the range of species that can be predicted, input parameters, how miRNAs and mRNAs are inputted, the ability to predict for novel miRNAs and to provide probability of miRNA-target interaction, and the number and format of outputted targets. 'Impact' relates to factors that reflect popularity of a given method, such as the number of the reviews that considered and highlighted a given predictor and their citation rates. The empirical evaluation concerns inclusion of an empirical study that compares several predictors. The empirical evaluation component lists the size of the test datasets, the inclusion of native non-functional pairs/true negatives (TN) and the number of evaluated methods. The year listed in brackets next to the number of evaluated methods is the year of the publication of the latest predictor that was empirically evaluated.

	Analytical description						Empirical evaluation		
	# of methods	Latest method	Described methodologies	Described evaluation procedures	Described availability	Assessed ease of use /impact	Dataset size	Inclusion of TN	# of evaluated methods (year newest)
This review (2014)	**38**	**2013**	**x**	**x**	**x**	**x**	**~100 thousands**	**x**	**7 (2012)**
Ref.[117] (2014)	10	2010	x		x	o			
Ref. [118] (2012)	20	2011	x		x				
Ref. [51] (2012)	11	2007	x						
Ref. [119] (2011)	11	2007	x						
Ref. [120] (2010)	7	2009	x		x	o			
Ref. [121] (2010)	30	2008	x		x	o			
Ref. [122] (2010)	10	2008	x		x	o			
Ref. [116] (2009)	10	2007	x		x		150	x	7 (2007)
Ref. [123] (2009)	8	2008	x						
Ref. [124] (2009)	14	2008	x						
Ref. [125] (2009)	9	2008	x		x				
Ref. [126] (2007)	9	2006	x		x	o			
Ref. [127] (2007)	14	2006	x		x				
Ref. [128] (2007)	9	2006	x		x				
Ref. [115] (2006)	20	2006	x		x	o	84		5 (2005)
Ref. [129] (2006)	10	2005	x		x	o			
Ref. [130] (2006)	11	2006			x	o	113	x	9 (2005)

4.2 **Overview of our review**

The systematic nature of our review stems from two facts. We define the reasons for our review and the list of methods that will be used to perform the review and to address these reasons. These methods were selected based on an explicit (documented here) search strategy that covers as much of the relevant literature as possible. We also make sure that we collect comprehensive and consistent information for each method, including the criteria to evaluate them empirically.

The reasons for our review include lack of existing, comprehensive, and up-to-date reviews (which is discussed in Section 4.1), lack of well-defined comparative assessments (which would use the same datasets, protocols and evaluation measures), and lack of contributions that provide practical guidance for the end users.

Our systematic review includes as many miRNA target prediction methods as possible. We used "miRNA target prediction" as the keywords to search all related articles in PubMed. Pubmed is the main source of biomedical literature from MEDLINE, life science journals and online books. We assume that all high-quality relevant works are available in this repository. We manually separated these articles into reviews and methods. We combined the resulting set of methods with all methods mentioned in the existing reviews (Table 4-1). We excluded articles for predictions in plants, which is motivated by an observation that predictions of targets in plants are relatively easy and are considered a solved problem [49, 131]. We also excluded meta predictors and methods that depend on experimental data (gene expression or CLIP-based data). To sum up, our inclusion criteria include sources from existing reviews and sources available in PubMed based on our query, while our exclusion criteria are focus on plants, meta predictors, and use of experimental data.

As a result, we review 38 predictors of miRNA targets in animals including recent methods that were not included in prior methods. We provide a comprehensive and practical summary of miRNA target prediction field for developers to design better prediction methods and for end users to select more appropriate tools. We address following questions:

1. What predictive models and features were utilized by the current predictors?
2. How did the current predictors take advantage of experimental validated miRNA targets and how did they evaluate their methods?

3. How easy to use are the current predictors for both typical users who have little knowledge of computing and advanced users? What is the impact of these methods with respect to applications and their influence to develop future prediction methods?
4. How good is the predictive quality of the current predictors? What factors should be considered when selecting a method from this large pool of available tools?

The first two and the last two questions target interests of the developers and the end users, respectively.

Taken together, we provide analysis from all key perspectives that are relevant to the end users and developers including overview of the mechanistic basis of miRNA-mRNA interaction and how this information is incorporated into the underlying predictive methodologies. We also give detailed summary of evaluation, usability and popularity/impact of the 38 predictors. As one often omitted dimension, we discuss the scope of the outputs, i.e., whether a given method provides a propensity score (probability of binding) or only a binary outcome (binding vs. non-binding), and whether it predicts positions of the miRNA binding sites on the target gene. We are the first to conduct an empirical comparative assessment on both low-throughput and high-throughput experimental data for the predictions at the miRNA-mRNA duplex (to predict whether a given fragment on mRNA interacts with a given miRNA) and gene (to predict whether a given mRNA interacts with a given miRNA) levels. We use four benchmark datasets and consider seven representative methods including recent predictors. We systematically evaluate both binary and (for the first time) real-valued propensity to compare multiple methods. Moreover, we utilize our in-depth analytical and empirical review to provide practical insights for the end users and developers.

4.3 **Datasets**

We developed four benchmark datasets using the miRTarBase repository, gene expression data from Gene Expression Omnibus (GEO) and pulsed stable isotope labeling by/with amino acids in cell culture (pSILAC). The miRTarBase introduced in section 2.3 provides the largest number of positive (functional/binding) and negative (non-functional/non-binding) miRNA-mRNA complexes. GEO is the largest source of microarray, sequencing and other forms of high-throughput genomic data [132]. pSILAC is a technique for quantitative proteomics [133]. Our

datasets cover human and mouse, which is motivated by research interests in using miRNAs in human-health related applications [134, 135] and our objective to include the largest possible number of predictors, i.e., relatively few methods work on other species.

We note that we do not consider positive and negative data in specific tissues. Although miRNAs were shown to be tissue-specific [136-138], studies suggest that expression levels of mRNA genes are not related to tissue types [139, 140]. This implies that tissue-specific miRNA-mRNA interaction relies on presence of the miRNA in a given tissue. As long as the miRNA can interact with the target mRNA, the interaction should happen if the miRNA is present in the corresponding tissue. Thus, our datasets are used to build predictive models and quantify their predictive performance when considering relationship between pairs of miRNA and mRNA without consideration of the tissue types. The corresponding results should be translated into specific tissues by screening for miRNAs that are present in these tissues.

The first dataset, called TEST_duplex, is used to assess the target site predictions at the duplex level. We selected targets that were validated by at least one of the low-throughput experimental methods which are considered as strong evidence: real-time polymerase chain reaction (qPCR), luciferase assay or western blot. We focused on targets that were released recently to limit overlap between our benchmark data and data used to develop the evaluated predictors. The functional targets deposited to miRTarBase after 2012 and all non-functional duplexes from human and mouse were included; we used all non-functional data because of their small number. The second, TEST_gene dataset focuses on the evaluation at the gene level. We selected miRNAs that have both functional and non-functional genes in miRTarBase and for which the functional genes were validated after 2012.

Furthermore, we extend our evaluation to analyze whether the current methods are capable of predicting at the transcriptome/proteome scale (to predict all possible mRNAs that interact with a given miRNA) using two additional datasets that rely on the annotations from the high-throughput methods. TEST_geo dataset is based on results from three microarray-based experiments: GSE6838, GSE7864 and GSE8501. The interactions for 25 miRNAs were annotated by contrasting expression arrays before miRNA transfection and at 24h after miRNA mimics were transfected [141-143]. As recommended in [144, 145], we remove genes for which the expression magnitudes are below the median in the control transfection experiments.

TEST_psilac dataset was originally developed in a proteomic study that utilized pSILAC technique [133, 146]. Previous studies assume that genes that are more repressed (characterized by higher drop in the expression levels) are more likely to be targeted by the transfected miRNA. These studies use a certain fraction of the genes with the highest magnitude of the decrease in the expression levels (repressed genes) as functional and the same fraction of the genes for which expression levels have increased by the largest margin (over-expressed genes) as non-functional [144, 147]. Instead of using an arbitrary fraction value to define the functional and non-functional pairs, we vary this value between 1% and 50%. Detailed summary of the four datasets is shown in the Table 4-2.

Table 4-2 Summary of the four benchmark datasets

Dataset	# of miRNAs	# of genes	# of functional miRNA-mRNA pairs	# of non-functional miRNA-mRNA pairs	Type of evaluation	Source of data
TEST_duplex	129	166	158	36	Gene sites	Low-throughput biochemical arrays
TEST_gene	45	221	150	115	Genes	Low-throughput biochemical arrays
TEST_geo	25	16097	109963	117911	Genes	High-throughput microarrays (GSE6838, GSE7864, GSE8501)
TEST_psilac	5	22327	19351	16525	Proteins	High-throughput pSILAC (http://psilac.mdc-berlin.de/download/)

The comprehensiveness of our tests stems from the fact that we consider targets as gene segments (TEST_duplex dataset), genes (TEST_gene and TEST_geo datasets) and proteins (TEST_psilac dataset). We also utilize different source of information that is used to perform annotations including low-throughput assays (TEST_duplex and TEST_gene datasets), microarrays (TEST_geo dataset) and pSILAC (TEST_psilac dataset).

4.4 Analytical description of the current miRNA target predictors

We consider 38 sequence-based methods, from the earliest predictor that was published in 2003 to the latest method that was released in 2013; chronological list of methods is shown in Table 4-3. We exclude the meta methods (since they are inconvenient to use and require availability of results from base methods) and approaches that rely on the experimental data. Most of the miRNA target predictors were developed by different research groups, with several groups that

continue maintaining and updating their algorithms. Cohen's group at EMBL proposed the first miRNA target predictor in 2003 [32] and updated it in 2005 [148]. TargetScan and TargetScanS were developed by Bartel at MIT and Burge at Cambridge [23, 33, 143, 149]. Another popular tool, DIANA-microT, which was created by Hatzigeorgiou group, has been recently updated to version 5.0 [109, 150-152]. Rajewsky's lab published their predictor PicTar in 2005 and updated it in 2011 [72, 153].

4.4.1 Predictive methodologies and mechanistic basis of miRNA-mRNA interaction

Table 4-3 summarizes types of predictive models and the underlying details of the miRNA-mRNA interactions that they utilize to predict miRNA targets. There are two categories of predictive models: heuristic and empirical. The heuristic models use screening algorithms that search positions along the mRNA sequences and scoring functions that filter targets by combining values of several inputs in an *ad-hoc* manner. Early predictors applied heuristic approaches owing to the lack of sufficient amount of data to build the empirical, knowledge-based models. Even today the scoring function-based designs are dominant (19 out of 38 methods) because of their easy setup, flexibility to integrate different types of inputs, and computational efficiency. The empirical models are inferred from a training dataset. Given the success of machine learning-based models in bioinformatics [154, 155] and growing size of the experimental data, since 2006 progressively more predictors utilize empirical machine learning models including SVMs, decisions trees, and artificial neural networks (ANNs).

Table 4-3 Methodologies and the corresponding mechanistic basis of miRNA-mRNA interaction used by the miRNA target predictors

We summarize key aspects including model type, region that is searched to predict targets, and inclusion of several mechanistic properties that are known to provide useful inputs for prediction, such as complementarity between miRNA and mRNA, site accessibility, and conservation across species; × means that a given aspect was irrelevant or not considered. Predictors are sorted in the chronological order. "Model type" describes type of predictive model type including screening of the mRNA sequence, heuristic scoring function, and empirically-designed genetic programming (GP), support vector machine (SVM), decision stump (DS), and artificial neural network (ANN) models. "Complementarity" indicates positions for which complementarity of WC base pairs between miRNA and mRNA was explored in the seed (first 8 positions on the miRNA) and non-seed regions. Four common seed types are 6mer, 7mer-A1, 7mer-m8 and 8mer; they have consecutive complementary WC base pairs on these positions. "1-8", "2-8", etc. annotations mean that these do not have to be consecutive complementary WC base pairs. Non-seed denotes the center and 3' end of the miRNA region where e.g., 38 nt means the size of the targets is up to 38 nucleotides; 14-20 nt indicate the non-seed regions is considered from the 14^{th} to 20^{th} nucleotide; "remaining" refers the region from the end of the seed to the end of the miRNA. "Site accessibility" describes inclusion of two aspects: AU content around the targets and free energy. If free energy is used then the name of the package used to calculate it is given (if known), otherwise ✓ is used. "Conservation" indicates species that were used in calculation of conservation: anopheles (a), chicken (c), drosophila (d), fungi (f), dog (g), human (h), mouse (m), nematode (n), rat (r), zebra fish (z), primate (P), mammal (M), and vertebrate (V). If conservation is used but species are unknown then ✓ is used. Methods that consider prediction of multiple sites on the same gene are annotated with ✓ in the "multiple sites" column. For machine learning methods, the "features" column indicates number of used features and whether and what feature selection approach was used; ✓ denotes the features are used but the count is unknown.

Predictor	Reference	Year published	Model type	Complementarity		Site accessibility		Conservation	Multiple sites	Features	
				seed	non-seed	free energy	AU %			count	selection
Stark *et al.*	Ref. [32]	2003	screening	1-8	miRNA size+5	mFold	×	a, d	✓	×	×
TargetScan	Ref. [33]	2003	score	7mer-m8	to 1^{st} mismatch	Vienna RNA	✓	m, r, z	×	×	×
DIANA-microT	Ref. [109]	2004	score	×	38 nt	✓	×	m	✓	×	×
RNAhybrid	Ref. [71]	2004	score	6mer	×	RNAhybrid	×	a, d	✓	×	×
miRanda	Ref. [156]	2004	score	7mer-m8	×	Vienna RNA	×	f, m, r	✓	×	×
Rajewsky's	Ref. [114]	2004	score	1-8	×	mFold	×	d	×	×	×
TargetScanS	Ref. [23]	2005	score	6mer	×	×	×	c, g, h, m, r	×	×	×
Robins	Ref. [157]	2005	score	2-8	×	Vienna RNA	×	×	×	×	×
Xie *et al.*	Ref. [24]	2005	score	8mer	×	×	×	g, h, m, r	✓	×	×
PicTar	Ref. [72]	2005	score	7mer-A1, 7mer-m8	remaining	mFold	×	d	×	×	×
MovingTarget	Ref. [158]	2005	screening	1-8	50 nt	DINAMelt	×	d	✓	×	×
MicroInspector	Ref. [159]	2005	score	7mer-A1, 7mer-m8	×	Vienna RNA	×	×	×	×	×
TargetBoost	Ref. [160]	2005	GP	pattern	30 nt	mFold	×	×	×	✓	×
Stark *et al.*	Ref. [148]	2005	score	6mer	10^{tn} nt to end	RNAhybrid	×	d	✓	×	×
miTarget	Ref. [161]	2006	SVM	2-7	20 nt	Vienna RNA	✓	×	×	15	wrapper
RNA22	Ref. [27]	2006	score	×	pattern	×	×	×	✓	×	×
MicroTar	Ref. [55]	2006	score	7mer-A1, 7mer-m8	×	Vienna RNA	×	×	✓	×	×
EIMMo	Ref. [162]	2007	Bayesian	7mer-A1, 7mer-m8	×	×	×	✓	×	✓	×
STarMir	Ref. [163]	2007	score	×	miRNA size	sFold	×	×	×	×	×
PITA	Ref. [164]	2007	score	6mer	×	Vienna RNA	×	×	✓	×	×
TargetRank	Ref. [165]	2007	score	6mer	×	×	✓	✓	✓	×	×
MirTarget2	Ref. [147]	2008	SVM	6mer	×	Vienna RNA	×	×	×	6	filter
HuMiTar	Ref. [79]	2008	score	6mer	9-13, 14-20 nt	×	×	×	×	×	×
TargetMiner	Ref. [78]	2009	SVM	6mer	13-16 nt	✓	✓	✓	✓	30	filter
TargetSpy	Ref. [77]	2010	DS	×	all	Vienna RNA	×	×	×	7	filter
Mtar	Ref. [76]	2010	ANN	6mer	remaining	Vienna RNA	×	×	×	16	×
mirSVR	Ref. [144]	2010	score	2-7	×	miRNAbind	✓	✓	×	×	×
SVMicrO	Ref. [75]	2010	SVM	5 patterns	remaining	Vienna RNA	✓	✓	×	39	wrapper
RepTar	Ref. [110]	2010	screening	6mer	remaining	Vienna RNA	×	✓	✓	×	×
PACMIT	Ref. [74]	2011	screening	×	remaining	Vienna RNA	×	×	✓	×	×
MultiMiTar	Ref. [166]	2011	SVM	6mer	13-16 nt	×	✓	×	✓	39	filter
miREE	Ref. [167]	2011	SVM	1-8	13-16nt, remaing	Vienna RNA	✓	×	×	25	filter
miRcode	Ref. [168]	2012	screening	7mer-A1, 7mer-m8	×	×	×	P, M, other V	✓	×	×
miRmap	Ref. [145]	2012	regression	6mer	remaining	Vienna RNA	✓	M	✓	12	filter
HomoTarget	Ref. [169]	2012	ANN	1-8	remaining	✓	×	×	×	12	filter
SuperMirTar	Ref. [170]	2013	Graph	6mer	12-17 nt	RNAhybrid	✓	×	×	×	×
Fujiwara's	Ref. [171]	2013	Cis-element	×	×	×	×	×	×	×	×
MIRZA	Ref. [172]	2013	Bayesian	1-8	remaining	✓	×	×	×	×	×

The predictive models use inputs that are derived from the knowledge of mechanistic details of the miRNA-mRNA interactions. The most commonly used predictive input is the complementarity of the WC base pairing between miRNAs and mRNAs. In contrast to the near-perfect WC base pairing in plants [49], animal miRNAs usually bind mRNAs with only some positions that are paired [50]. Complementarity of the WC base pairing in the seed region is particularly important; only six methods did not treat the seed differently from the non-seed region. To compare, 15 methods did not consider complementarity in the non-seed region. The major types of complementarity in the seed include 6mer, 7mer-A1, 7mer-m8, and 8mer (section 2.2.1). Some methods consider binding of the first eight nucleotides as important but do not restrict it to particular seed types. Moreover, several predictors (HuMiTar [79], TargetMiner [78], MultiMiTar [166], miREE [167], and SuperMirTar [170]) also suggest specific positions which are more useful for the prediction. These methods, except for HuMiTar, use machine learning models and empirical feature selection to find these positions. One other exception is that TargetBoost [160], RNA22 [27] and SVMicrO [75] utilize patterns of complementarity generated from native miRNA-mRNA complexes, rather than focusing on the seed types.

The site accessibility and evolutionary conservation inputs are used to increase specificity. The accessibility is relevant since miRNA-mRNA interaction requires binding of a relatively large RISC [51]. This input is quantified with content of adenine and uracil nucleotides (AU content) and free energy that estimates stability of the mRNA sequences. Most target predictors employ existing software, like Vienna RNA package [106], mFold [173], DINAMelt [174] and sFold [175], to calculate the free energy. Authors of RNAhybrid claim that their own approach prevents intra-molecular base pairing and bulge loops, which leads to improved estimates of the free energy [71]; this approach was also used in the predictor by Stark et al. [148] and in SuperMirTar [170]. Most predictors calculate the free energy of the miRNA-target duplexes. However, several methods (MicroTar [55], STarMir [164], PITA [163], TargetMiner [78], SVMicrO [75], PACMIT [74], and miREE [167]) calculate arguably more relevant relative energy which is the hybridization energy lost by miRNA-mRNA binding minus the disruption energy needed for opening up the local mRNA structure of the target. Several studies found that enriched AU content in mRNA 3' UTRs is important for the interaction with miRNAs [53, 54, 176]. This was exploited in 2003 in TargetScan, even before experimental data that verifies the effect was published [143]. Since then several methods have used this information (see "AU %"

column in Table 4-3). Use of the evolutionary conservation of miRNA targets is motivated by a premise that "similar" species should share common miRNAs and their targets. However, this leads to omission of the non-conserved targets [55, 56]. The value of the inclusion of the target conservation remains an open question; Table 4-3 reveals that conservation is used less frequently in recent years. Still, methods that search for targets in long coding DNA segments (CDSs) use conservation to improve specificity [152, 168, 177, 178]. Based on an observation that targeting of multiple sites enhances the mRNA regulation [57, 58], 17 out of the 38 methods increase the propensity of binding to a target gene with multiple predicted sites (see "Multiple sites" column in Table 4-3).

The machine learning models often use empirical approaches to select inputs (features) that are relevant to the prediction of miRNA targets. Table 4-3 shows that the count of the selected features ranges from a few to a few dozen; these features quantify specific aspects related to the complementarity, accessibility, and conservation. The considered feature selection approaches include wrapper- and filter-based methods. The former approach searches for the best subset of features to maximize predictive performance of a given machine learning model. Filters rank features according to a metric, like F-score or correlation, and select a predefined number of the top-ranked features.

4.4.2 Evaluation protocols

Benchmark datasets used to develop and test the predictors and the corresponding evaluation procedures are summarized in Table 4-4. Many early methods were designed/evaluated using data only from *Drosophila* due to limited availability of validated miRNA targets in other species. However, even some early predictors (TargetScan (33), DIANA-microT [33], miRanda [179] and TargetScanS [23]) considered higher eukaryotes. More recent methods generally cover more species. Interestingly, in 14 cases predictors were validated on test datasets but there was no mention about data being used to design these predictive models. This may mean that the test data was used in the design, e.g., to set thresholds and parameters. HuMiTar was the first method that was properly tested on an independent (from the training set) dataset [79]. Even with the currently available relatively large number of validated miRNA targets, only a few recent predictors (TargetMiner [78], TargetSpy [77], Mtar [76], MultiMiTar [166] and miREE [167]) were trained and tested on different (independent) datasets. Moreover, the sizes of some training

Table 4-4 Protocols for evaluation of the miRNA target predictors

We describe the benchmark datasets used to design and test the predictors including the target "species", size of training and test datasets, and source of the non-functional samples. × means that a given aspect was irrelevant or not considered. The "species" are anopheles (a), chicken (c), drosophila (d), fungi (f), dog (g), human (h), mouse (m), nematode (n), rat (r), virus (v), zebra fish (z), mammals (M) and vertebrates (V). "# training/test duplexes" is the number of functional (+) and non-functional (-) samples if they were provided; otherwise ✓is used. The "non-functional samples" describes the sources of the non-functional examples; they include targets with validated lack of interaction with a given miRNA or artificially generated (via shuffling or randomization) samples. We also describe procedures used to assess the predictive performance of the predictors. This includes the number of the experimentally validated targets, criteria used to measure the performance, and whether statistical tests and functional analysis were performed. "# of validated targets" shows the number of experimentally tested predicted targets. The "criteria" lists the criteria used to assess the programs: signal-to-noise ratio (*SNR*), false positive rate (*FPR*), area under *ROC* curve (*AUC*), Matthews correlation coefficient (*MCC*) and average class-wise accuracy (*ACA*). Methods for which predictions were assessed with statistical tests of significance and for which functional analysis was performed are indicated with ✓ in the "statistical test" and "functional analysis" columns, respectively.

Predictor	Reference	Benchmark datasets				Evaluation procedures			
		species	# of training duplexes	# of test duplexes	non-functional samples	# validated targets	criteria	statistical test	functional analysis
Stark *et al.*	Ref. [32]	d	×	5+	shuffled miRNA	6	*SNR*, conservation	✓	✓
TargetScan	Ref. [33]	h m p	×	gene level	shuffled miRNA	11	*FPR*, *SNR*	×	✓
DIANA-microT	Ref. [109]	h	×	11+	shuffled miRNA	0	*SNR*	×	×
RNAhybrid	Ref. [71]	d	×	11+	shuffled miRNA	0	*SNR*	✓	×
miRanda	Ref. [156]	h z	×	8+	shuffled miRNA	0	*FPR*	×	×
Rajewsky's	Ref. [114]	d	25	gene level	random mRNA	0	*FPR*	×	×
TargetScanS	Ref. [23]	V	×	×	shuffled miRNA	0	*SNR*	×	✓
Robins	Ref. [157]	d	×	×	×	10	×	✓	×
Xie *et al.*	Ref. [24]	h	×	×	×	12	×	×	×
PicTar	Ref. [72]	d	×	19+	shuffled miRNA	0	*SNR*, sensitivity	×	✓
MovingTarget	Ref. [158]	d	×	×	×	3	×	×	×
MicroInspector	Ref. [159]	d	×	×	×	0	×	×	×
TargetBoost	Ref. [160]	d n	36+, 3000-	×	random mRNA	0	*AUC*	✓	×
EMBL	Ref. [148]	d	×	gene level	shuffled miRNA	8	×	✓	✓
miTarget[147]	Ref. [161]	h	152+, 246-	same with training	4-mer on non-positives	0	*AUC*	×	✓
RNA22	Ref. [27]	d h n m	×	21+	shuffled miRNA	168	*FPR*	×	×
MicroTar	Ref. [55]	d m n	×	63, 13 and 43+	×	0	sensitivity	×	×
EIMMo	Ref. [162]	d n z M	×	120 in all	validated	0	sensitivity, specificity	×	×
STarMir	Ref. [163]	d n	×	39+, 12-	validated	0	*FPR*, *SNR*	×	×
PITA	Ref. [164]	d	×	123+, 67-	validated	0	*AUC*	×	×
TargetRank	Ref. [165]	V	×	×	×	0	×	×	×
MirTarget2	Ref. [147]	c g h m r	✓	✓	validated	0	*AUC*	×	×
HuMiTar	Ref. [79]	h	66 in all	39 and 190 in all	validated	0	*AUC*, *SNR*	×	×
TargetMiner	Ref. [78]	h	289+, 100-	187+, 59-	microarray+validated	0	*MCC*, *ACA*	×	×
TargetSpy	Ref. [77]	c d h m r	3872+, 4540-	61+, 59-/102+, 88-	pSILAC+validated	0	*AUC*	×	×
Mtar	Ref. [76]	h	150+, 200-	190+, 200-	validated	0	*AUC*	×	×
mirSVR	Ref. [144]	h	gene level	gene level	microarray+CLIP	0	*AUC*	×	×
SVMicrO	Ref. [75]	h m r	324+, 3492-	gene level	microarray	0	*AUC*	×	×
RepTar	Ref. [110]	h m v	197 and 22 in all	same with training	validated	0	precision, accuracy	✓	×
PACMIT	Ref. [74]	d h	137+, 83-/2406+, 13400-	same with training	pSILAC+validated	0	specificity and *pROC*	✓	×
MultiMiTar	Ref. [166]	h	289+, 289-	187+, 57-	pSILAC+validated	0	*MCC*, *ACA*	×	×
miREE	Ref. [167]	d h m n r v z	324+, 351	2 new datasets	pSILAC+PAR-CLIP+validated	0	*pROC*	✓	×
miRcode	Ref. [168]	V	×	×	×	0	×	×	✓
miRmap	Ref. [145]	h m	gene level	same with training	Microarray; CLIP	0	×	×	×
HomoTarget	Ref. [169]	h	112 pos + 313 neg	same with training	validated	0	*AUC*	×	×
SuperMirTar	Ref. [170]	h m	2860 human, 582 mouse	674+, 15132-	pSILAC+valiated	0	*AUC*	×	×
Fujiwara's	Ref. [171]	h	×	155+	validated	0	*pROC*	×	×
MIRZA	Ref. [172]	all available	gene level	same with training	Ago2-CLIP	0	sensitivity	×	×

datasets are relatively small (a few dozen samples) and some datasets are unbalanced and have more artificial non-functional samples than the functional samples; some datasets use only a few validated non-functional samples. A particularly challenging aspect is a low number of experimentally validated non-functional samples, i.e., an mRNA validated not to interact with a given miRNA. Several early methods utilized artificial non-functional data created by either shuffling miRNA sequences or by randomization of mRNAs; these approaches were criticized to generate unrealistic samples [78]. More recent attempts scan the mRNA transcripts where validated target sites or Ago-binding sites are masked and use the target segments with at least 4-mer matches in the seed region or one mismatch or G:U wobble in the 6-mer seed as the non-functional samples [76, 144, 161]. This approach assumes that the knowledge of functional targets or Ago-binding sites is complete, while in fact these computationally generated non-functional miRNA-mRNA pairs could be functional. Some recent methods label over-expressed genes when particular miRNA mimics are added to cells as non-functional, but data from this limited number of miRNAs may be biased. These various attempts to generate the benchmark datasets may result in mislabeling, over-fitting the training datasets, and unrealistic (possibly inflated) evaluation of predictive performance.

We also analyze the evaluation procedures. The early predictors were evaluated primarily based on signal-to-noise ratio (*SNR*) between the number of predicted targets in functional genes and in true or artificial non-functional genes. PicTar was the first to report sensitivity, based on only 19 native targets. TargetBoost and miTarget were the first to utilize more informative *ROC* curves, but with the caveat of using artificial non-functional data. The criteria used to evaluate predictive quality vary widely between methods. Some measures are biased by the composition of the dataset (e.g., accuracy and precision) and provide incomplete picture (e.g., sensitivity without specificity and vice versa). This makes comparisons across predictors virtually impossible. The standards to compare between methods are also relatively low, as in most cases evaluation did not include statistical tests. On the positive side, the assessment of several methods included experimental validation of targets. The authors of RNA22 method performed a large-scale validation and claimed that 168 out of 226 tested targets were repressed; however, they did not found whether these targets were bound by the specific miRNAs. Some primarily older methods also included functional analysis of the predicted targets.

4.4.3 Usability and impact

Table 4-5 shows that miRNA target predictors are available to the end users as webservers, standalone packages, pre-computed datasets, and upon request. The 21 methods that are provided as webservers are convenient for *ad-hoc* (occasional) users. The 13 standalone packages are suitable for users who anticipate a high-throughput use and/or who would like to include them into their local software platforms; most of them are also available as the webservers. The convenience of access to pre-computed results is provided for 10 methods. However, these predictions may not be updated timely and do not include results for novel miRNAs that are continually generated.

The ease of use is affected by the use and number of parameters, scope of predictions, format of inputs, and ability to predict targets for novel miRNAs. The prediction methods rely on parameters that can be used to control how prediction is performed, e.g., the seed size, the number of allowed GU wobbles and mismatches, selection of mRNA regions that are searched, and the cut-offs for free energy and predicted propensity score. These parameters are usually set based on experience of the designer or user of a given method, or are optimized empirically using a dataset. Eleven methods hardcode and hide these parameters from the users, which arguably makes them easier to use but also reduces ability of the end users to tune the models for specific needs or projects. RNAhybrid [71] offers eight (the most) parameters for tuning; RepTar and PITA [110, 164] have seven and five parameters, respectively; and eight predictors allow adjusting between one and four parameters. Importantly, these predictors provide default values for the parameters, so they can be seamlessly used even by layman users.

A "user-friendly" method should allow predicting a wide range of species and target types. Most of the early methods only allow predictions in the 3'UTRs, except for RNAhybrid [71], miRanda [180], DIANA-microT-CDS [152] and PACMIT-CDS [178] that also search coding DNA sequences (CDSs) and TargetScanS [23] and Xie's method [24] that consider open reading frames (ORFs) and promoters, respectively. As more miRNA targets were discovered beyond the 3'UTRs [177, 181], several newer programs (RNA22 [27], STarMir [163], Mtar [76] and miRcode [168]) predict in the 3'UTRs, CDSs, and 5'UTRs. A few methods (RNAhybrid [71], MicroInspector [159], MicroTar [55] and MIRZA [38]) do not limit species for which they predict. They accept target genes as RNA sequences or provide standalone packages where users

Table 4-5 Usability and impact of the miRNA target predictors

We summarize availability, ease of use, and impact/popularity. × means that a given aspect was missing. ~ denotes unknown as the information was not available in the paper or in the webserver. "Availability" focuses on type of implementation available to the end user: standalone (s), webserver (ws), pre-computed results (p) and upon request (ur), and provides the corresponding URLs. The links shown in shade did not work. "Ease of use" covers aspects related to the scope of a given method and ease to run it including the number of input parameters of the corresponding webservers, the targets regions and species that can be predicted, the approximate number of predicted targets, the format in which the searched genes are provided and the ability to predict for new miRNAs. "Target region" indicates where a given method searches for targets: untranslated region (UTR), coding DNA segment (CDS), and open reading frame (ORF). The covered species are chicken (c), drosophila (d), chimpanzee (e), dog (g), human (h), mouse (m), nematode (n), opossum (o), rat (r), cow (w), thale cress (t), zebra fish (z), and vertebrate (V). The estimated count of predicted targets per miRNA per gene, or per miRNA only (for predictors do not allow inputting target gene) which is denoted by *, is given in the "# of targets" column; counts were estimated based on the corresponding papers or by testing the webservers. The possible formats of the input genes are by name, by sequence, or by either name or sequence; "none" denotes that searching particular genes is not allowed. "new miRNA" shows whether a given method allows to predict new miRNAs.; methods that allow inputting miRNA sequences can be used to predict new miRNAs and are annotated with ✓; otherwise ×. "Impact/popularity" is assessed using the number of times a given method was highlighted and considered in the 15 review papers listed in Table 4-2; "# citations" lists the average count of citations per year since published collected in Sept. 2013 using the ISI Web of Knowledge.

Predictor	Availability		Ease of use						Impact/popularity		
	type	URL	# para-meters	target region	covered species	# targets	format of gene	new miRNA	high-lighted	consi-dered	# citations
Stark *et al.*	×	×	~	3'UTR	d	~	~	×	0	4	34.4
TargetScan	s ws	http://www.targetscan.org/	3	3'UTR	d h m n z	a few	name	✓	3	14	429.5
DIANA-microT	ws	http:http://diana.pcbi.upenn.edu/DIANA-microT	1	3'UTR	d h m n r t	a few	name	✓	2	14	38.4
RNAhybrid	s ws	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	8	3'UTR, CDS	any	dozens	sequence	✓	2	12	69.3
miRanda	s ws p	http://www.microrna.org/microrna/home.do	0	3'UTR, CDS	d h m r n	1000s*	none	×	0	15	104.8
Rajewsky's	×	×	~	3'UTR	d	~	~	~	0	2	18.5
TargetScanS	s ws	http://genes.mit.edu/tscan/targetscanS2005.html	0	3'UTR, ORFs	d m n other V	100s*	name	×	4	10	429.5
Robins	×	×	~	3'UTR	~	~	~	~	0	2	14.6
Xie *et al.*	×	×	~	promoters and 3'UTR	~	~	~	~	0	2	124.8
PicTar	ws p	http://pictar.mdc-berlin.de/	0	3'UTR	d h m n	dozens	name	×	1	16	26.9
MovingTarget	ur	×	~	3'UTR	~	~	~	~	0	5	5.7
MicroInspector	ws	http://bioinfo.uni-plovdiv.bg/microinspector/	2	3'UTR	any	a few	either	✓	0	3	13.0
TargetBoost	demo	http://www.interagon.com/demos.html	~	3'UTR	n	~	~	~	0	8	7.9
Stark *et al.*	p	http://mirnas.russelllab.org/	~	3'UTR	~	100s*	~	~	1	5	67.8
miTarget	×	http://cbit.snu.ac.kr/miTarget	~	3'UTR	~	~	~	~	0	9	10.4
RNA22	ws p	https://cm.jefferson.edu/rna22v1.0/	4	3'UTR, CDS	d h m n	1000s*	name	×	0	13	80.8
MicroTar	s	http://tiger.dbs.nus.edu.sg/microtar/	0	3'UTR	any	~	~	✓	0	3	5.6
EIMMo	ws p	http://www.mirz.unibas.ch/ElMMo3/	0	3'UTR	d h m r n z	a few	name	×	1	6	17.4
STarMir	ws	http://sfold.wadsworth.org/cgi-bin/starmir.pl	0	3'UTR, CDS, 5'UTR	h m	dozens	either	✓	0	1	28.3
PITA	s ws	http://genie.weizmann.ac.il/pubs/mir07/index.html	5	3'UTR	d h m n	a few	either	✓	1	6	97.6
TargetRank	ws	http://hollywood.mit.edu/targetrank/	0	3'UTR	h m	100s*	none	×	0	1	22.9
MirTarget2	ws p	http://mirdb.org/miRDB/	0	3'UTR	c g h m r	a few	name	×	0	4	26.4
HuMiTar	ur	×	3	3'UTR	h	~	sequence	✓	0	1	2.2
TargetMiner	ws p	http://www.isical.ac.in/~bioinfo_miu/targetminer20.htm	0	3'UTR	h	a few	name	×	0	2	8.7
TargetSpy	s ws	http://www.targetspy.org/	2	3'UTR	c d h m r	a few	name	×	0	1	7.8
Mtar	×	×	~	3'UTR, CDS, 5'UTR	~	~	~	~	0	1	4.0
mirSVR	×	×	~	3'UTR	~	~	~	~	0	1	52.0
SVMicrO	s	http://compgenomics.utsa.edu/svmicro.html	~	3'UTR	h m r	~	~	~	0	1	4.3
RepTar	s ws p	http://bioinformatics.ekmd.huji.ac.il/reptar/	7	3'UTR	h m	a few	name	×	0	0	2.5
PACMIT	×	×	~	3'UTR	~	~	~	~	0	0	7.3
MultiMiTar	s ws	http://www.isical.ac.in/~bioinfo_miu/multimitar.htm	0	3'UTR	h	a few	either	✓	0	0	3.3
miREE	s	http://didattica-online.polito.it/eda/miREE/	0	3'UTR	d h m n r z	dozens	either	✓	0	0	1.0
miRcode	ws p	http://www.mircode.org	3	3'UTR	h	a few	name	×	0	1	5.3
miRmap	s ws p	http://mirmap.ezlab.org/	4	3'UTR, CDS, 5'UTR	c e h m o r w z	a few	either	✓	0	0	4.0
HomeTarget	×	×	~	3'UTR	~	~	~	~	0	0	0.5
SuperMirTar	×	×	~	3'UTR	~	~	~	~	0	0	0.0
Fujiwara's	ur	×	~	3'UTR	~	~	~	~	0	0	0.0
MIRZA	s ws	http://www.clipz.unibas.ch/index.php?r=tools/sub/mirza	0	3'UTR	any	s few	sequence	✓	0	0	14.0

can prepare their own mRNA database. Most of the other predictors are constrained to human, mouse, fly and worm. The latter two were the first two species that were used to study miRNA targets. Seven methods consider a more restrictive set of species including human and mouse, and four of them also predict for rat or chicken. Four recent methods (HuMiTar [79], TargetMiner [78], MultiMiTar [166] and miRcode [168]) focus on human mRNAs, and TargetBoost [160] works only in worms. Besides, miRanda [180] is the only method in our review that provides expression levels of miRNAs in different tissues.

Next, we analyze format of the inputs. The target genes can be specified by the name or identifier, by the mRNA sequence, or are preloaded and the user is not allowed to enter them. Entering the name (e.g., GenBank Accession, NCBI gene ID and/or name) is arguably convenient but it also limits the prediction to the mRNAs that are available in the considered reference database(s). Allowing the user to provide mRNA sequence alleviates this drawback. Six predictors (MicroInspector [159], STarMir [163], PITA [164], MultiMiTar [166], miREE [167] and miRmap [145]) accept either the name or the sequence, while three and eleven programs accept only sequences or names, respectively. The miRNAs can be inputted in two formats: by name and/or by sequence. Again, although it may be convenient to specify miRNAs by their names, this is a rather substantial drawback which does not allow predicting for novel miRNAs that are nowadays discovered at a rapid pace. Six methods that offer webservers (TargetScan [33], DIANA-microT [152], MicroInspector [159], PITA [164], miREE [167] and miRmap [145]) accept either the miRNA name or the sequence, while three and ten only take the sequences or the names, respectively. Table 4-5 reveals that twelve methods can predict targets of novel miRNAs.

When considering the outputs, the number of predicted targets varies widely between methods. Table 4-5 reports that while most methods predict a few targets per gene per miRNA, some predict hundreds, while miRanda [180] generates hundreds of thousands of targets per miRNA.

One way to measure impact/popularity of a given method is to analyze its inclusion in prior reviews. Considering the 16 reviews (Table 4-2), 29 out of the 38 methods were included in at least one review and 11 in five or more. Moreover, five reviews highlighted/recommended certain predictors. TargetScan [33] and TargetScanS [23] were recommended in 3 and 4 reviews, respectively; DIANA-microT [109] and RNAhybrid [71] twice, and EMBL method [148],

PicTar [72], EIMMo [162] and PITA [163] once. We also calculated the average citation counts per year since a given predictor was proposed, using the Web of Knowledge. Table 4-5 reveals that 21 of the 38 methods receive on average over 10 citations per year and all methods published before 2008 receive at least five citations per year. Three early methods receive over 100 citations every year. TargetScan/TargetScanS [23] is on the extreme end (400+ citations per year), and this could be attributed to its popularity and convenient availability, the fact that empirical studies often compare to this predictor, and since it is widely used in practical applications.

4.5 Empirical comparison of selected miRNA target predictors

We selected several representative predictors for the empirical evaluation. The selected methods have to be conveniently accessible to the end users via a webserver or a pre-computed database. They also have to cover human and mouse, predict target sites (to perform evaluation at the duplex level), and provide propensity (probability) of the interaction. Using these filters we selected eight methods (see Table 4-6). We use the latest versions of these methods, except for PicTar2 which is substantially different from PicTar and no longer qualifies as a sequence-based predictor. PicTar 2005 was first published in 2005; five methods including TargetScan 6.2, miRanda 2010, EIMMo3, miREE and mirTarget2 v4 were proposed or updated between 2010 and 2012; and two in 2013: DIANA-microT-CDS and miRmap v1.1. We excluded miREE from the evaluation since this method did not predict any targets on our TEST_duplex and TEST_gene datasets. The remaining seven methods use a diverse set of predictive models, with four that utilize heuristic scoring functions and three that use the machine learning models including Bayesian classifier, SVM and regression. miRmap was built based on gene expression data, while the other methods were derived based on the low-throughput experimentally validated data.

We assess the original predictive models proposed in the publication of the selected prediction methods on our benchmark test datasets, which is a common practice in this field [115]. We did not rebuild their predictive models on a benchmark training dataset, like it is done in some other fields, because this is not a common practice and this could introduce a bias since different methods in this field were purposely developed using different datasets. Moreover, most of the

predictors are simulated by their pre-computed predictions that are available online and there is no code to retrain them. Thus, we collected predictions from these methods using either their online webservers or downloadable pre-computed predictions. We recorded their predicted binding targets (mRNA sequences and/or positions of the binding site on mRNA) and the corresponding propensities (real-valued scores that quantify probability of the miRNA-target interaction).

Table 4-6 Summary of the criteria used to select methods for the empirical assessment

The covered species are chicken (c), drosophila (d), dog (g), human (h), mouse (m), nematode (n), rat (r) and zebra fish (z). A selected method must at minimum predict for human and mouse. The format of input gene is by name, by sequence, either by sequence or name, or none in case when this input is hardcoded in a given method. The "outputs" summarize the format and scope of the outputs generated by a given method. The formats of outputs include real-valued propensity of the miRNA-mRNA interactions (probability or score) and binary outcome (binding vs. non-binding). A selected method must provide the more informative real-valued probability. The target sites can be tracked by the predicted position on the mRNA sequence, by the matching of the seed on the mRNA sequence, by both of these options, or the output does not allow the tracking. We rejected methods that do not allow the tracking since this is required to perform analysis at the miRNA-mRNA duplex level. The "notes" list extra features or important drawback of the predictors. Six out of seven evaluated predictors are capable of batch predictions, which facilitated our tests. miREE did not predict any targets in either TEST_duplex or TEST_gene datasets and thus was excluded. The features that resulted in rejection of a given method from the empirical evaluation are given in italic font on shaded background. Names of the selected methods are given in bold font.

miRNA target predictor	covered species	format of input gene	outputs		notes
			score	target site tracking	
TargetScan v6.2	c, d, h, m, z	name	probability	both	batch search
RNAhybrid v2.1	any	sequence	*binary*	sequence	*always predicts a target*
DIANA-microT-CDS	d, h, m, n, r	name	probability	both	batch search
miRanda v2010	d, h, m, n, r	none	probability	sequence	batch search
PicTar v2005	d, h, m, n	name	probability	seed sequence	
MicroInspector v1.5	any	either	*binary*	both	
RNA22 v2.0	d, h, m, n	name	*binary*	both	
PITA v2007	d, h, m, n	either	*binary*	position	
STarMir v2007	h, m	either	probability	position	*long runtime*
EIMMo3	d, h, m, n, r	name	probability	position	batch search
TargetRank v2007	h, m	none	probability	*none*	
MirTarget2 v4.0	c, g, h, m, r	name	probability	both	batch search
TargetMiner v2012	*h*	name	*binary*	seed sequence	
TargetSpy v1.0	c, d, h, m, r	name	*binary*	sequence	
RepTar v1.2	h, m	name	*binary*	sequence	
MultiMiTar	*h*	either	score	sequence	
miREE	c, h, m, n, r, z	name	probability	both	*did not predict targets*
miRcode v11	*h*	name	*none*	position	
miRmap v1.1	c, e, h, m, o, r, w, z	either	probability	both	batch search
MIRZA	any	*30-50nt long sequence*	probability	sequence	*input gene sequences are limited to between 30 and 50nt in length*

Table 4-7 and Table 4-8 summarize results of the assessment at the gene level on the TEST_gene dataset and the duplex level on the TEST_duplex dataset. A given miRNA-target pair was predicted as functional if the target was predicted using the corresponding miRNA; the remaining targets were assumed to be predicted as non-functional and the corresponding propensity was set to 0. When assessing the gene level predictions, we scored a given gene using the sum of propensities among all its predicted target sites for a given miRNA. Since these seven methods were initially published before 2012, we use experimentally validated miRNA targets that were published after 2012 to perform the empirical assessment. This limits a bias caused by a potential overlap between our benchmark data and data used to develop a given method.

Table 4-7 Comparison of predictive performance at the gene level (TEST_gene dataset) and at the duplex level (TEST_duplex dataset)

We evaluate seven representative targets predictors. We measure area under the *ROC* curve (*AUC*), Matthews Correlation Coefficient (*MCC*), sensitivity (Sen.), specifity (Spe.), precision (Prec.), signal-to-noise ratio for predicted functional (*SNR+*) and predicted non-functional targets (*SNR-*) and predicted-to-native functional target ratio (*PNR*). Methods are sorted in the descending order by their *AUC* values. The best value of each measurement across all the predictors is given in bold font.

Prediction type	Predictor	*AUC*	*MCC*	Sen.	Spe.	Prec.	*SNR+*	*SNR-*	*PNR*
At the duplex level	TargetScan	**0.674**	0.200	**0.823**	0.389	0.855	1.346	**2.194**	**0.962**
	DIANA-microT	0.673	**0.273**	0.627	0.722	**0.908**	**2.256**	1.934	0.690
	miRmap	0.658	0.158	0.741	0.444	0.854	1.333	1.713	0.867
	miRanda	0.560	0.081	0.437	0.667	0.852	1.310	1.184	0.513
	EIMMo	0.552	0.116	0.696	0.444	0.846	1.253	1.463	0.823
	PicTar	0.538	0.069	0.272	**0.806**	0.860	1.400	1.107	0.316
	MirTarget2	0.519	0.055	0.285	0.778	0.849	1.282	1.088	0.335
At the gene level	TargetScan	**0.748**	0.386	0.733	0.652	0.733	2.108	2.446	**1.000**
	EIMMo	0.725	**0.391**	0.707	0.687	0.746	2.257	2.342	0.947
	miRmap	0.714	0.353	**0.800**	0.539	0.694	1.736	**2.696**	1.153
	DIANA-microT	0.637	0.225	0.520	0.704	0.696	1.759	1.467	0.747
	miRanda	0.636	0.239	0.467	0.765	0.722	1.988	1.435	0.647
	MirTarget2	0.627	0.298	0.327	**0.922**	**0.845**	**4.174**	1.369	0.387
	PicTar	0.588	0.196	0.340	0.835	0.729	2.058	1.265	0.467

Considering the predictions of the miRNA-mRNA duplexes, TargetScan and DIANA-microT secure the highest *AUC* values of 0.674 and 0.673, respectively. Moreover, DIANA-microT has the highest *MCC*, which improves over the second best TargetScan by 0.073 (relative improvement of (0.273-0.200)/0.200*100%=36.8%). TargetScan offers the highest sensitivity, i.e., it correctly predicts the largest fraction of the functional duplexes. On the other hand, PicTar has the highest specificity, i.e., it correctly predicts the largest number of the non-functional duplexes. This means that functional targets predicted by PicTar are likely to be functional.

DIANA-microT offers the highest *SNR*+. TargetScan has the highest *SNR*-, relatively good *SNR*+, and very good *PNR*. *PNR* value of TargetScan reveals that it only slightly under-predicts, by 3.8%, the number of functional duplexes. The other methods, except for miRmap and EIMMo, under-predict the functional duplexes by a large margin. We illustrate relation between predictive quality (*SNR* values) and the outputted propensities binned to 10 intervals in Figure 4-2A. The number of predicted duplexes and their *SNR* values in each interval are denoted by size and color of the bubbles (dark blue for accurate predictions), respectively. Alternating red and blue bubbles for a given predictor indicate that values of its propensity do not correlate with the underlying predictive quality. All methods have blue bubbles for propensity of 0, which means that predict the non-functional duplexes well. However, predicted functional targets (propensity > 0) are often inaccurate (red bubbles with black borders) particularly for lower values of propensity. DIANA-microT predicts well when its propensity > 0.7, and miRmap and TargetScan when > 0.4 and 0.8, respectively. Analysis of statistical significance reveals that the differences in the *AUC* values (results above diagonal in Table 4-8) are not statistically significant between TargetScan, DIANA-microT, and miRmap. However, these three predictors are significantly better than the other four methods (p-value≤0.001).

Table 4-8 Statistical significance of the differences in predictive performance measured with *AUC* for predictions at the gene level (TEST_gene dataset) and at the duplex level (TEST_duplex dataset)

Results below (above) diagonal are for the predictions at the gene (duplex) level. The statistical tests are based on 10 repetitions of randomly chosen 50% of the duplexes/genes from the TEST_duplex or TEST_gene datasets. +/=/- indicate that the *AUC* value of a predictor in the corresponding column is significantly larger/not significantly different/significantly smaller than that of the method in the corresponding row. Methods are sorted in the descending order by the *AUC* values at the gene level. Several levels of p-values are used: "++" or "--" for p-value < 0.0001, "+" or "-" for 0.0001 < p-value ≤ 0.01, and "=" for |p-value|>0.01.

Duplex / Gene	TargetScan	miRmap	DIANA-microT	miRanda	EIMMo	PicTar	MirTarget2
TargetScan		--	=	=	--	--	-
miRmap	++		++	+	=	=	=
DIANA-microT	+	=		=	-	--	-
miRanda	++	++	++		-	--	-
EIMMo	++	++	+	=		=	=
PicTar	++	++	+	=	=		=
MirTarget2	++	++	++	+	+	=	

Table 4-9 analyzes anticipated predictive performance at the duplex level based on information that is available before the prediction is performed, including the nucleotide composition of the seed region and the overall size of the input miRNA sequences. The hints summarized in this

Table could guide selection of a predictor based on the miRNA sequences. Most methods, especially TargetScan, DIANA-microT and miRmap, predict well for medium-sized (22 nucleotides long) miRNAs. The predictions for longer miRNAs are generally less accurate. Considering the nucleotide content in the seed region, the same three methods provide high-quality predictions for miRNAs when the seeds have 2 adenines or 2 guanines, and <2 cytosines. DIANA-microT also predict well for <2 adenines and >2 uracil and miRmap for <2 adenines. Overall, we recommend TargetScan, DIANA-microT and miRmap since their *AUCs* > 0.7 for specific types of miRNAs.

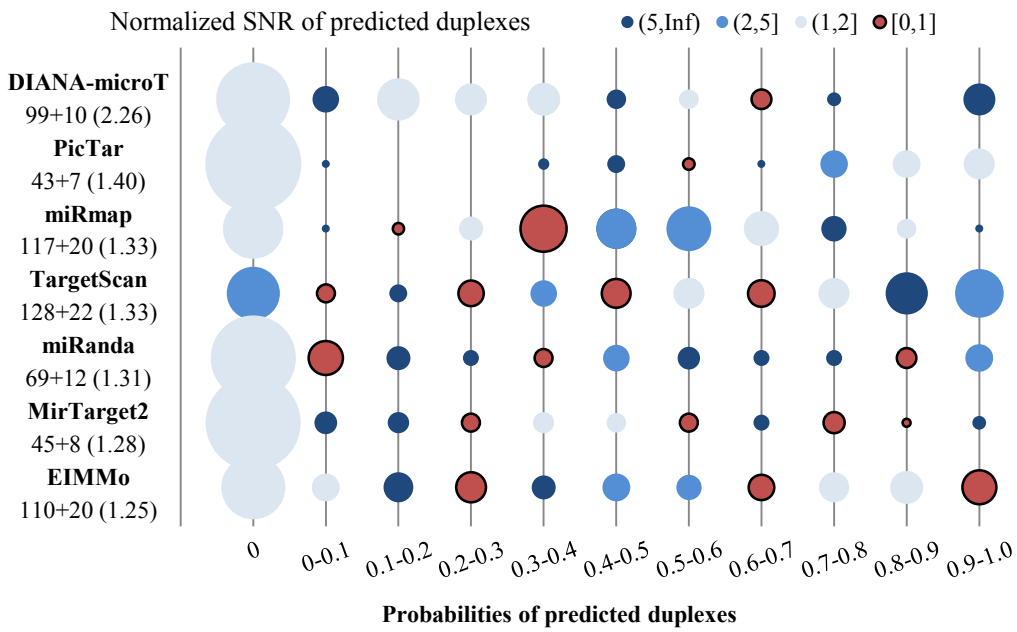
Table 4-9 Relation between predictive quality measured with *AUC* and compositional characteristics of the input miRNAs for predictions at the duplex level (TEST_duplex dataset)
The compositional characteristics include the size of miRNA and the count of each nucleotide type in the seed region. The sizes are divided into short (<22 nt), medium (=22 nt) and long (>22 nt). The count of nucleotides in the seeds of miRNAs is grouped into low (<2 nt), medium (=2 nt) and high (>2 nt). The *AUC* values obtained by a given predictor are coded as: "-" for [0, 0.55], "=" for (0.55, 0.6], "+" for (0.6, 0.7] and "++" for (0.7, 1.0].

Predictor	Size of miRNAs			A count			C count			G count			U count		
	short	medium	long	low	medium	high	low	medium	high	low	medium	high	low	medium	high
TargetScan	+	++	-	+	++	+	++	=	++	+	++	+	+	+	+
DIANA-microT	-	++	=	++	++	-	++	=	++	+	++	-	+	+	++
miRmap	=	++	-	++	++	=	++	-	++	+	++	+	+	+	+
miRanda	=	+	-	+	=	-	=	-	+	=	-	=	=	=	-
EIMMo	=	+	-	-	+	-	+	-	=	=	=	=	-	+	-
PicTar	-	=	-	-	-	=	=	-	=	-	-	+	=	=	-
MirTarget2	-	=	-	=	-	-	=	-	=	-	-	=	-	-	-

The overall prediction quality is higher and ranking of the methods is slightly different for the predictions on TEST_gene dataset when compared to the TEST_duplex dataset (Table 4-7). TargetScan secures the highest *AUC* while EIMMo moves up to the second place and provides the highest *MCC*. TargetScan improves in *AUC* over the second best EIMMo by 0.023 (relative improvement of 3.2%) and over miRmap by 0.043 (relative improvement of 4.8%). miRmap offers the highest sensitivity and TargetScan provides arguably the best balance between sensitivity and specificity (both scores are high and similar). MirTarget2 is the most conservative method given its highest specificity, precision and *SNR*+, i.e., it predicts only a few functional targets but with high success rate. The *PNR* values reveal that TargetScan predicts exactly the right number of functional genes and EIMMo only 5.3% too few. Figure 4-2B shows relation between predictive quality (*SNR* values) and the propensities generated by the prediction methods. Interestingly, predictions associated with higher propensities are more likely to be more

accurate, as evidenced by the presence of (dark) blue bubbles. As a highlight, EIMMo predicts well in every propensity bin, and the targets predicted by TargetScan and miRanda with propensities over 0.3 and 0.4, respectively, are characterized by high *SNR* values. Analysis of statistical significance of differences in the *AUC* values (results below diagonal in Table 4-8) reveals that TargetScan's results are significant better (p-value≤0.001) compared to the other predictors. *AUCs* of EIMMo and miRmap are not significantly different and significantly higher than *AUCs* of the other four methods (*p*-value≤0.001). We also analyze relation between predictive performance at the gene level and the number of target sites predicted in a given gene (Figure 4-2C). Most methods, except for MirTarget2 and miRanda, can predict three or more target sites per gene for a given miRNA. We observe that predictive quality for genes for which at least two sites are predicted is better (bubbles have darker blue color), particularly for EIMMo, TargetScan and miRanda. This suggests that for these predictors higher number of predicted sites could be used as a marker of higher predictive quality.

A



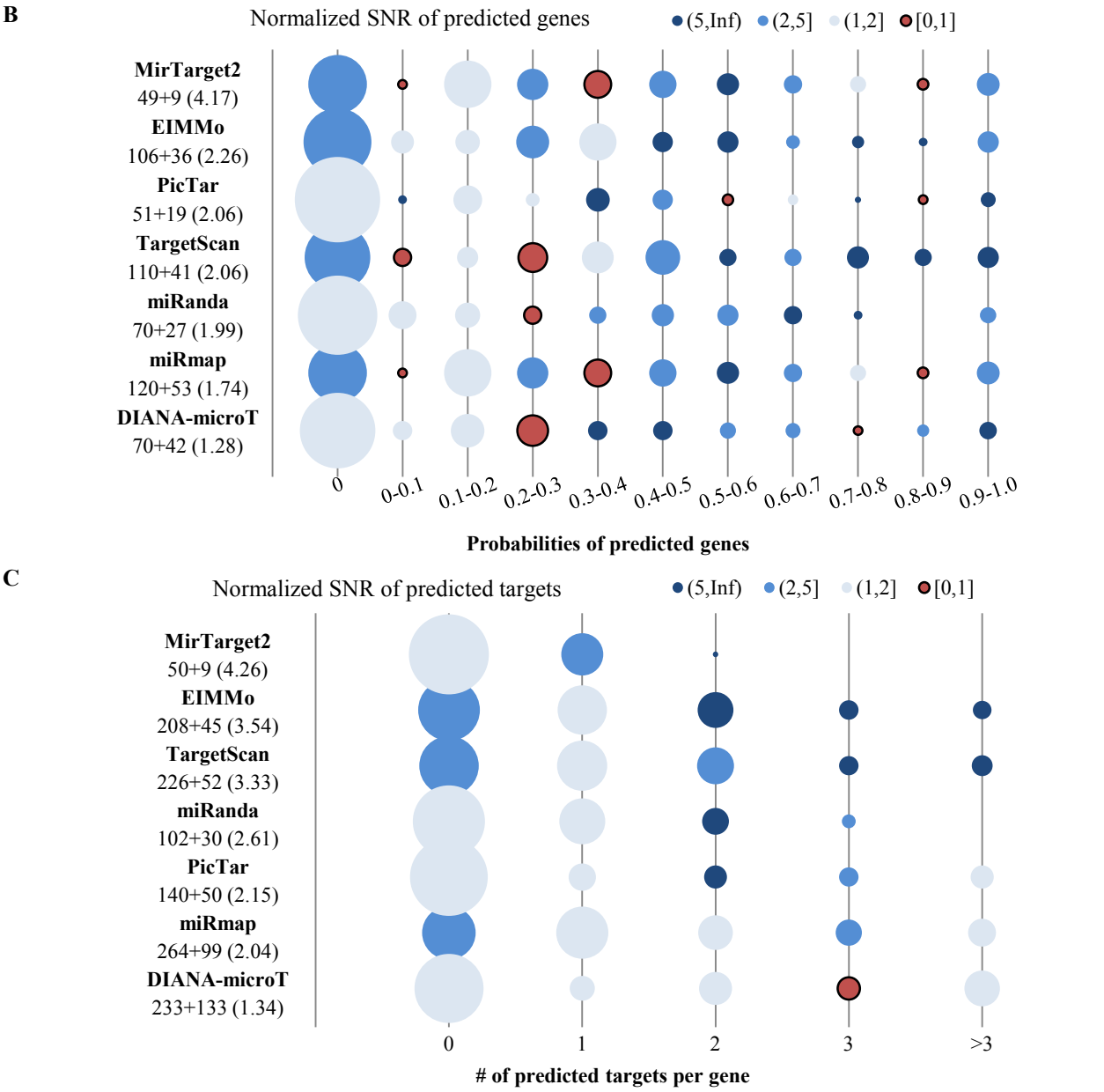
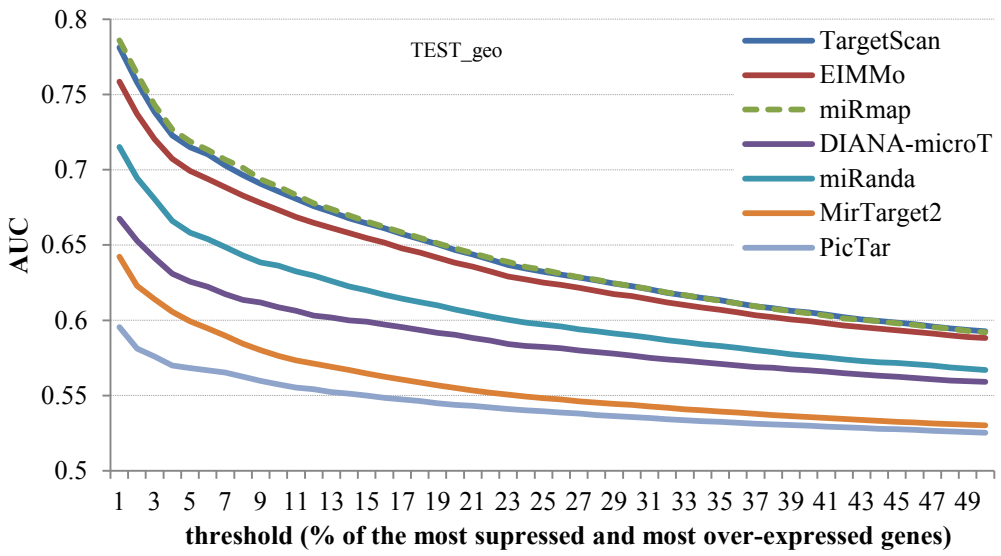


Figure 4-2 Relation between the predictive quality measured with normalized *SNR* and the predicted real valued probability of a given duplex or a miRNA-mRNA pair being functional on the TEST_duplex (panel A) and TEST_gene (panel B) dataset, and the number of predicted targets per gene (panel C)

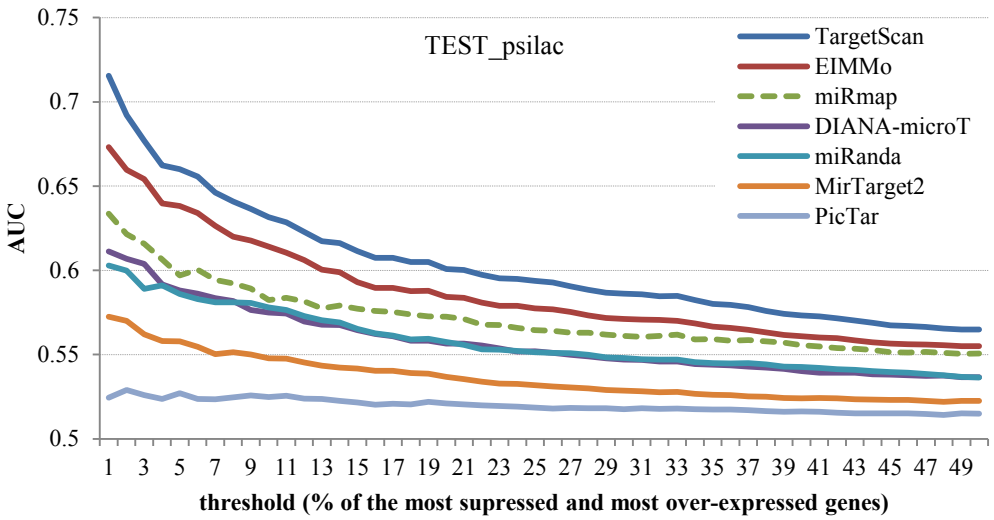
Probabilities are divided into 10 bins. The size and colors of the bubbles in the three panels denote the number of predicted targets/genes and the normalized *SNR* of targets/genes, respectively. Methods are sorted in the descending order by their overall *SNR+* values. The number of predicted functional and non-functional targets/genes and the corresponding normalized SNR values (shown in the brackets) are given below the name of a given method. The ratio of predicted functional targets/genes in true functional to non-functional genes is calculated for bubbles for probability (*x*-axis) >0, while the ratio of predicted non-functional targets/genes in true non-functional to functional genes is calculated for bubbles for probability (*x*-axis) =0. The raw data used to draw this figure comes from Table 4-8.

Predictions at the transcriptome/proteome scale on the TEST_geo and TEST_psilac datasets are evaluated at different thresholds that define the fraction of the most repressed and most over-expressed genes that are annotated as functional and non-functional, respectively (Figure 4-3). *AUCs* are generally higher at the gene level (TEST_geo dataset, Figure 4-3A) than at the protein level (TEST_psilac dataset, Figure 4-3B). Considering the three gene-level datasets, the ranking of the methods on the TEST_psilac dataset is the same as on the TEST_gene dataset, and slightly different on the TEST_geo dataset. Based on the microarray data, miRmap achieves the best *AUC* which is comparable with the *AUC* of TargetScan and EIMMo. These three predictors have *AUCs* >0.7 when evaluated on the top 4% of genes with largest expression changes; using this threshold on average each miRNA targets 176 mRNAs. We note miRmap was originally trained and tested on two of the three microarrays from the TEST_geo dataset, so its predictive quality on this dataset could be overestimated. Considering the pSILAC data, only TargetScan provides *AUC* >0.7 when using top 1% of proteins for which expression levels change most; this threshold results in an annotation where on average each miRNA regulates 39 proteins. Overall, the *AUC* values decrease when more ambiguous genes (genes for which expression changes are weaker) are included, i.e., the fraction of the included repressed and over-expressed genes is higher. Analysis of the *MCC* values (Figure 4-3C and Figure 4-3D) leads to similar conclusions. TargetScan, EIMMo and miRmap secure the highest values of this index.

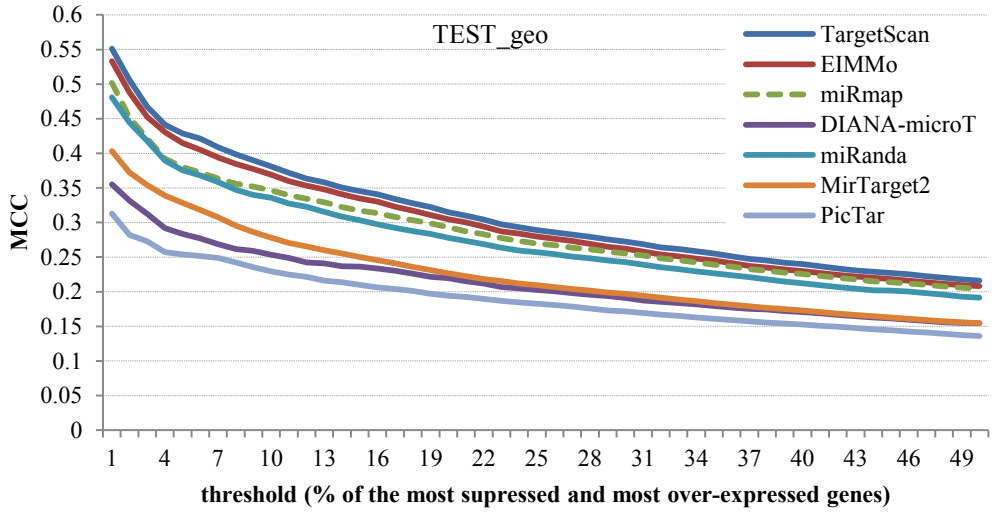
A



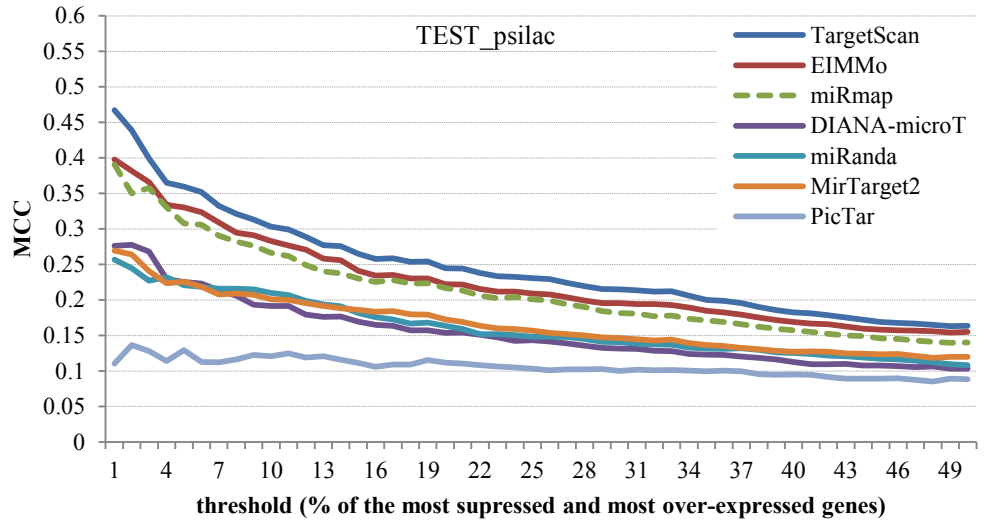
B


TEST_psilac
TargetScan
EIMMo
miRmap
DIANA-microT
miRanda
MirTarget2
PicTar
AUC
0.75
0.7
0.65
0.6
0.55
0.5
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
threshold (% of the most supressed and most over-expressed genes)

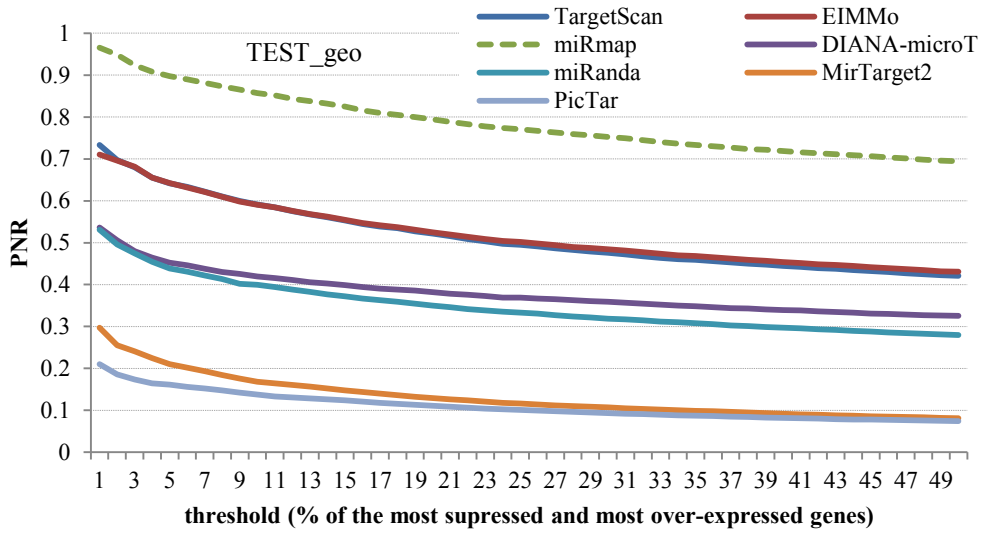
C


TEST_geo
TargetScan
EIMMo
miRmap
DIANA-microT
miRanda
MirTarget2
PicTar
MCC
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49
threshold (% of the most supressed and most over-expressed genes)

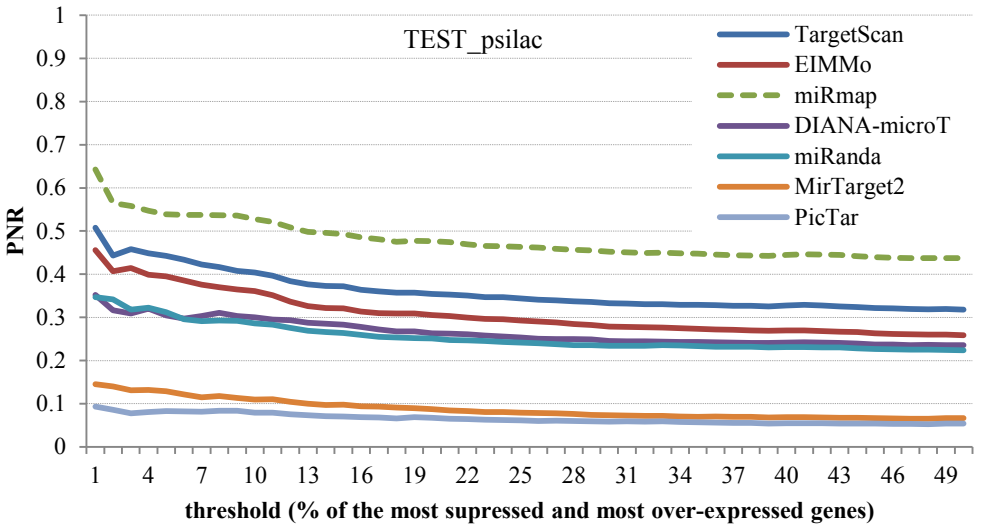
D


TEST_psilac
TargetScan
EIMMo
miRmap
DIANA-microT
miRanda
MirTarget2
PicTar
MCC
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49
threshold (% of the most supressed and most over-expressed genes)

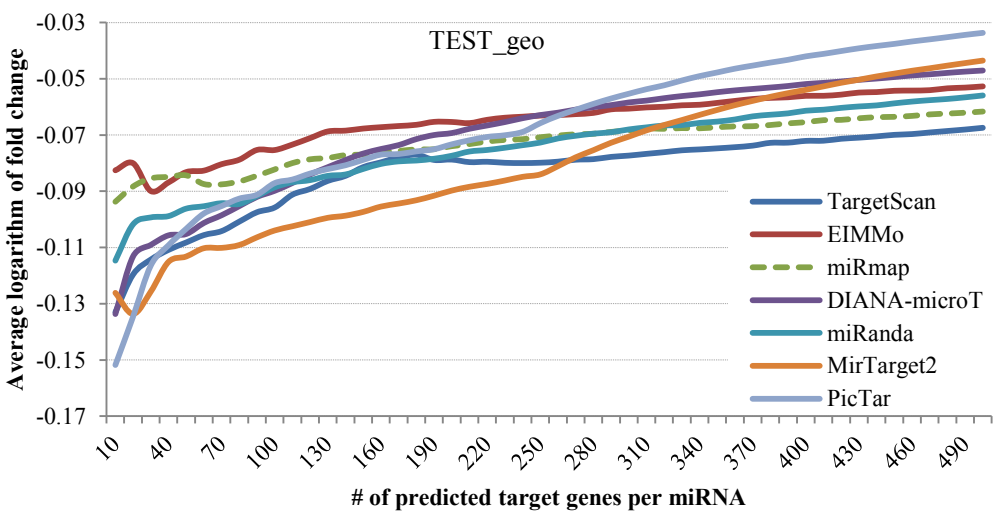
E


TEST_geo
TargetScan
EIMMo
miRmap
DIANA-microT
miRanda
MirTarget2
PicTar
PNR
threshold (% of the most supressed and most over-expressed genes)

F


TEST_psilac
TargetScan
EIMMo
miRmap
DIANA-microT
miRanda
MirTarget2
PicTar
PNR
threshold (% of the most supressed and most over-expressed genes)

G


TEST_geo
TargetScan
EIMMo
miRmap
DIANA-microT
miRanda
MirTarget2
PicTar
Average logarithm of fold change
# of predicted target genes per miRNA

H

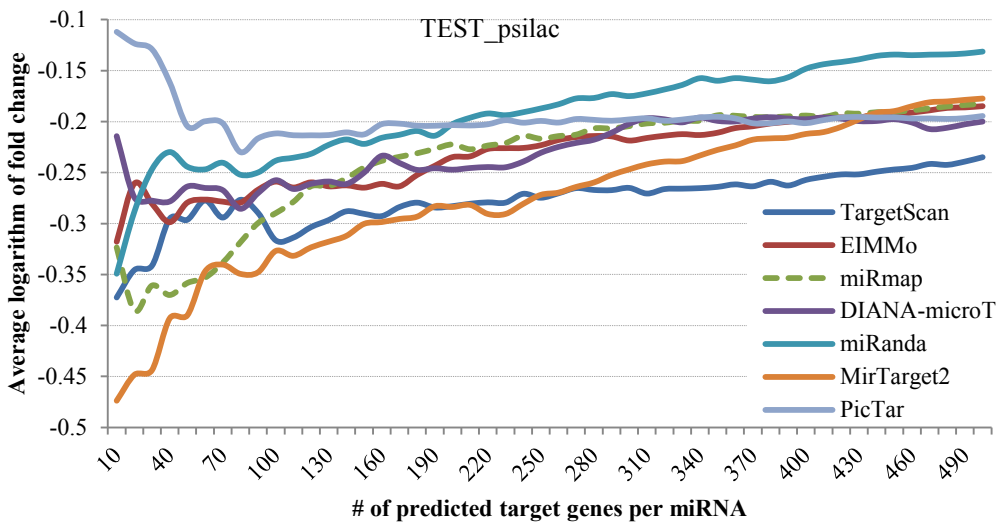


Figure 4-3 Relation between *AUC*, *MCC* and *PNR* values and the thresholds used to define the functional (most suppressed) and non-functional (most over-expressed) genes for the predictions on the TEST_geo (panel A, C and E) and TEST_psilac (panel B, D and F) datasets. Average logarithm of fold change of top predicted targets on the TEST_geo (panel G) and TEST_psilac (panel H) datasets

Methods are sorted in the same order with those on TEST_gene dataset. miRmap that was trained on the gene expression data is given with the dashed line.

We also calculate the average logarithm of the fold change of the top predicted genes (i.e., genes that obtain the highest propensity score) for each method to assess whether higher propensity implies better predictive performance (Figure 4-3G and Figure 4-3H). Genes with high propensity of binding predicted by MirTarget2 are characterized by large expression changes, with almost 3-fold change for the top 10 targets predicted for each miRNA. This strong result is consistent with high precision at the gene level on the TEST_gene dataset which is secured by this method. High values of propensities generated by TargetScan are also indicative of higher changes in the gene expression levels, while the results of the other methods are inconsistent between the two datasets. We note that expression level changes are larger on the TEST_psilac dataset, which is probably due to a different amount of mRNAs and available miRNAs in the cell [182] and differences in the experimental conditions. This also hints that it would not be plausible to predict absolute gene expression changes solely based on the miRNA and mRNA sequences. From the *PNR* curves (Figure 4-3E and Figure 4-3F), we observe that all methods, except for miRmap on the TEST_geo dataset, under-predict functional targets by a substantial margin. Considering that this datasets may miss native functional genes that are associated with smaller expression level changes and that some of the targets genes could be annotated based on

an indirect interaction with the miRNAs, the list of functional targets defined solely by the expression changes could be incomplete and may include false positives. Therefore, we do not expect *PNR* values close to 1 on the TEST_geo and TEST_psilac datasets.

4.6 **Conclusions**

We reviewed 38 miRNA target predictors from all significant perspectives including their prediction models, availability, impact, user friendliness, and protocols and measures that are used to evaluate their predictive performance. We found that standardized evaluation procedures are urgently needed since currently predictors are evaluated using different measures, different test protocols, and using vastly different datasets. This hinders comparison among these methods and appropriate selection by the end users. To this end, we empirically and systematically compared seven representative predictors on four benchmark datasets, considering prediction of miRNA-mRNA duplexes and target genes and proteins.

We found that although certain methods, like TargetScan and miRmap, offer high overall predictive quality, there is no universally best predictor. For instance, PicTar and MirTarget2 provide predictions with high specificity and low number of false positives (incorrectly predicted functional genes/duplexes). Thus, these two methods are suitable for users that would like to obtain a small subset of accurately predicted functional duplexes or genes. EIMMo predicts very well at the gene level. We observe that the count of functional target sites or genes predicted by TargetScan is the closest to the native count (*PNR* value close to 1), and thus this method should be used to accurately estimate the number of miRNA targets. We found that genes predicted as functional based on a higher number of sites are more likely to be accurate, particularly for the EIMMo and TargetScan predictors. Finally, the benchmark datasets and empirical results that we provide are useful to develop and comparatively assess future prediction methods.

We observe that predictions at the duplex level are characterized by lower predictive quality than the predictions of target genes. This agrees with intuition that predicting target sites, which require locating the right binding positions on the correctly predicted target gene, should be more difficult than predicting target genes only. Moreover, our estimates of the predictive performance are often lower than the estimates from the original publications. Possible reasons are:

1. We use experimental validated data which is likely more challenging than the artificial data that were used to assess previous predictors.
2. The non-functional validated duplexes that we use have relatively many WC base pairs in the seed regions (83% have at least 6 pairs). These sites were likely hypothesized to be functional, refuted and thus annotated as non-functional. This is why they have such seeds, which in turn makes them more challenging to separate from the functional duplexes when compared to a more "random" site.
3. MiRanda, PicTar, EIMMo and MirTarget2 provide only pre-computed predictions which may not include most up-to-date miRNA and transcript databases.

Unfortunately, we could not compare results with the previous reviews [115, 116, 130], because they did not consider a balanced selection of measurements (e.g., only provided sensitivity and precision which ignore true negatives), and such one-sided evaluation would not be meaningful.

Our review offers in-depth insights that could be used by the end users to select prediction methods based on their predictive performance (Table 4-7) and their input miRNAs (Table 4-9). We also provide several practical observations that consider specifics of applications of interest. Arguably, the commonly considered characteristics of the applications of the miRNA target predictors include the need to consider novel miRNAs and to focus on certain regions in the mRNAs, to predict a more complete or smaller and more accurate list of targets, to predict for a large set of miRNAs, to tweak desired parameters of the miRNA-mRNA interaction, and to generate propensities for the predicted interactions. We address these characteristics as follows:

- Only some methods can predict targets for novel miRNAs (see "new miRNA" column in Table 4-5).
- Applications that focus on particular regions (e.g., 5' UTR, CDS, promoters) should utilize predictors that were designed to consider these regions (see "target region" column in Table 4-5).
- Some methods generate few and potentially more accurate targets while some predict a larger and more complete set of targets that may include more false positives (see "# targets" column in Table 4-5). Users should choose an appropriate method depending on whether they look for a more complete or a more accurate set of targets.

- When predicting for a large number of miRNAs, the downloadable pre-computed results or methods that provide application program interfaces (APIs) should be used (see "batch search" in the "note" column in the Table 4-6).
- The end users should apply predictors with tunable seed type parameter, such as PITA, when searching for targets that utilize a particular seed type. Also, when aiming to find targets with low number of WC pairs in the seed region, only some predictors that consider such targets, like miREE, can be used.
- When predicting the target sites, the methods that can only predict target genes cannot be used (see "target site tracking" column in Table 4-6).
- Only some predictors provide predictions with the associated propensities of the interaction; many methods only provide binary (functional vs. non-functional) predictions (see "score" column in Table 4-6)

Chapter 5

Development of an Accurate and Novel *Ab Initio* Predictor of Non-canonical MicroRNA Targets

We reviewed 38 current miRNA target predictors in the previous chapter. Those methods vary in the predictive methodologies they use, usage, predictive performance, and other aspects. In this chapter we analyze potential drawbacks of the current predictors and introduce a new method that addresses these drawbacks.

5.1 Motivation

The predictions of virtually all methods depend on the complementary Watson-Crick (WC) base pairing in the seed region [33]. The common seed types are 6mer, 7 mer-A1, 7mer-m8 and 8mer, which are defined in section 2.2.1. The current predictors usually utilize the seed type as an input and feed it into their pre-set scoring functions or learning models together with other inputs. The design of these predictive methodologies relies on an assertion that the more WC base pairs are found in the seed, the more likely it is that the given miRNA interacts with the corresponding mRNA. Consequently, mRNA sites with more matches for a given miRNA are predicted more often than mRNA sites with fewer matches. We find a relatively large number of predicted 8mer and 7mer targets compared to the number of predicted 6mer targets that were retrieved by the current and popular methods. More specifically, MultiMiTar includes only 15.6% predicted 6mer targets in its pre-computed database [166] while TargetScan [183], MirTarget2 [147] and TargetSpy [77] are at the extreme end since they do not even predict the 6mer targets. Therefore, we define the 7mer and 8mer targets that are covered by the current methods as canonical targets, and the other (6mer or fewer) as non-canonical targets.

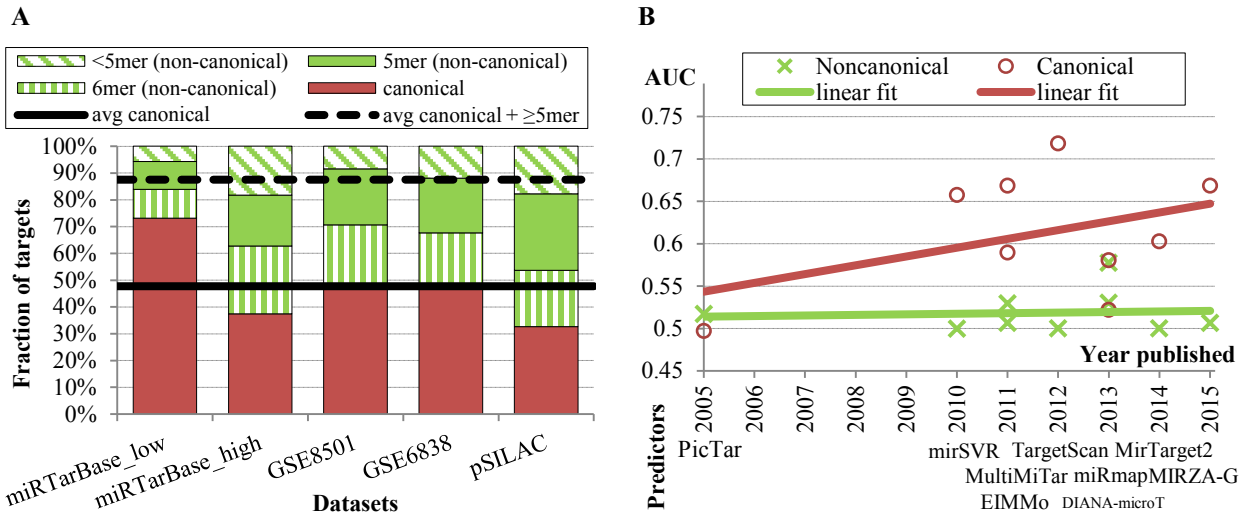


Figure 5-1 Fraction of canonical and non-canonical miRNA targets and predictive quality of current miRNA target predictors

(A) Results are reported on five datasets: (1) miRTarBase_low and (2) miRTarBase_high include data validated by low-throughput biochemical and high-throughput sequencing experiments from the miRTarBase database; (3) GSE8501 and (4) GSE6838 are two microarrays datasets; and (5) pSILAC annotates miRNA-mRNA interactions based on quantitative proteomics. We use top 50% of the most repressed genes for the latter three datasets. The fractions of canonical and non-canonical targets were calculated for each dataset as follows: 1) we use screening algorithm (the same as used by TargetScan) which slides the "seed" of the input miRNA over the 3' UTR of mRNA transcript to fold miRNA and mRNA segments into duplexes; 2) For each pair of miRNA and mRNA, we use the interaction site(s) with the maximal number of WC base pairs in the seed to annotate it as either canonical (8mer and 7mer) or non-canonical (6mer, 5mer and <5mer); and 3) we count the fraction of canonical targets and non-canonical targets over all miRNA:mRNA pairs. Solid (dashed) horizontal lines show the average fraction of canonical targets (fraction of combined canonical and non-canonical targets that exclude targets with less than 5 matches (<5mer targets)) computed based on the fractions of canonical targets in the five datasets. (B) *AUC* values of nine popular predictors are assessed on the canonical and non-canonical targets from the benchmark dataset (see section 5.3). The methods are sorted chronologically by year of publication and their names are given at the bottom. The linear fit into *AUC* values of the nine methods is shown using solid lines.

A number of studies have shown that interactions of miRNAs with the non-canonical targets are functional [34-36]. We comprehensively investigated the abundance of the non-canonical miRNA targets in animals across a wide range of data sources where targets were validated by different experimental methods including low-throughput biochemical assays, high-throughput microarrays, RNA-seq and pSILAC. We generated five datasets including over 50 thousands targets of more than 700 human and mouse miRNAs that we collected from miRTarBase repository [60], GEO [132] and pSILAC [146] based data. We found that on average only 47.7% of the miRNA targets in the 3' UTR of mRNAs are canonical, which is estimated by the average fraction of canonical targets across the five datasets (miRTarBase_low, miRTarBase_high, GSE8501, GSE6838 and pSILAC datasets) (Figure 5-1A), with 73% in the miRTarBase data

that was verified with low-throughput biochemical assays (miRTarBase_low dataset), 37% in the miRTarBase's data based on high-throughput sequencing (miRTarBase_high dataset), below 48% in two sets of microarrays (GSE8501 and GSE6838 datasets) and 33% in the pSILAC dataset. Our result confirms observations from earlier studies which found that the fraction of the non-canonical targets varies between 25% and 85% depending on the type of the high-throughput experiments used [37-39]. The annotations in the miRTarBase_low dataset include a large amount of canonical targets likely because they were handpicked among the favorable canonical seed types. Importantly, many of the current predictors require a minimal number of matches in the seed region to work. Some popular methods, such as TargetScan [23], MirTarget2 [147], miRcode [168], only search canonical targets which means that they cannot accommodate the other half of the targets. We empirically evaluated predictions of the newest versions of nine popular methods separately for the canonical and non-canonical targets using our benchmark TEST_gene dataset (see section 5.3) (Figure 5-1B). The predictive quality for the predictions of the canonical targets has improved over the last decade (red linear fit line in Figure 5-1B). Multiple more recent methods secure relatively good predictive performance ($AUC > 0.65$). However, the prediction of the non-canonical targets is characterized by poor predictive quality with *AUCs* at around 0.5 (green linear fit line in Figure 5-1B), which is equivalent to a random prediction. The main reason is that current predictors were not designed to identify this category of targets because of the lack of awareness of the high abundance of the non-canonical targets.

Besides these empirical results, we summarize the ability of 21 current and convenient for the end user predictors, defined as those that provide a working webserver or pre-computed database of their results, to predict non-canonical miRNA targets (Table 5-4). We divided these methods into three groups: 1) methods that cannot predict non-canonical targets; 2) predictors that can find the non-canonical targets but which were never evaluated for these predictions; and 3) methods for which predictions of the non-canonical targets were evaluated but which were not designed to predict these targets. None of the current methods was designed specifically for predicting non-canonical miRNA targets. The first and second groups include five and thirteen methods, respectively. Although some of them can predict non-canonical targets (usually limited to 6mer targets), their corresponding predictive quality is relatively low; Figure 5-1B shows results for a subset of 7 representative methods from these two groups. The last group includes three methods and for completeness we extended it by another method, miRTCat [184], that was

also evaluated but which is not "end user convenient" as its webserver is no longer maintained. Therefore, to date, only four methodology articles acknowledged this problem and evaluated the predictions for the hard-to-predict non-canonical targets. However, these methods were not designed to address this issue.

MIRZA [185] was published two years ago and relies on the availability of the CLIP (cross-linking immunoprecipitation) data, which substantially limits its applications, and predicts only a small subset of at most 25% of targets (including both canonical and non-canonical targets). The other three methods, miRanda-mirSVR [144], miRTCat [184] and MIRZA-G [186] require only the miRNA and mRNA sequences (sequence-based methods) and relax the restriction on the binding in the seed to allow for finding the non-canonical targets. However, miRTCat provides predictions that are limited to just 33 miRNA families, and the other two methods, miRanda-mirSVR and MIRZA-G, correctly predict only 5% and 2% of the non-canonical targets (Figure 5-1B), respectively.

5.2 Overview of proposed solution

Motivated by the high levels of abundance of the non-canonical targets and the fact that the existing methods either rely on the experimental data or find a very limited subset of these targets, we report a new sequence-based model that is designed for high-throughput non-canonical Mirna Target prediction (ncMirTar). Similar to the significant majority of current predictors (30 out of 38 [111]) we focus on the prediction in the 3' UTR of mRNA; four methods search the whole mRNA transcript including CDS and 5' UTR (RNA22 [27], STarMir [163], Mtar [76] and miRcode [168]), and four consider 3' UTR and CDS (DIANA-microT-CDS [152], RNAhybrid [71], miRanda [144] and PACMIT [74]). ncMirTar works in three steps. In the first step, the input 3' UTR mRNA sequences are scanned against a given miRNA and divided into two subsets: canonical targets that have at least one canonical site and non-canonical targets for which all sites are non-canonical. A list of the canonical targets is generated for the prediction with one of the existing miRNA targets predictors, and the non-canonical targets are predicted in the following two steps. In the second step, the input miRNA and 3' UTR mRNA sequences are converted into a small empirically designed set of numerical features. In the third step, these features are inputted into one of two machine learning models that are designed for the prediction of the 6mer and 5mer targets, depending on the duplex predicted for the given miRNA

and mRNA in the first step. This way ncMirTar complements predictions of the current methods for the canonical targets with its own predictions for the non-canonical targets. The key design aspects of ncMirTar are:

1. We utilize carefully selected, collected from multiple reliable sources, and non-canonical target-oriented training dataset that includes over 400 targets for 127 miRNAs to empirically design our method.
2. Our design seamlessly combines two SVM-based predictive models that capture characteristics of the two main types of non-canonical miRNA targets (5mer and 6mer targets account for about 76% of the non-canonical targets, see Figure 5-1A).
3. Each of the two models utilizes a handful of empirically selected features that quantify essential information from the input miRNA and mRNA sequences. This includes information used by other predictors, such as the complementarity of WC base pairing in a predicted miRNA-mRNA duplex and accessibility and conservation of the mRNA site, which we empirically combine together to optimize predictive performance. We also include a novel feature based on ranking of the seed type in the putative miRNA-mRNA duplex and we empirically demonstrate that its inclusion provides a significant boost to the predictive performance, allowing for accurate prediction of the non-canonical targets.

5.3 **Datasets**

MicroRNA targets that we use were validated by diverse sets of methods including low-throughput biochemical methods and high-throughput sequencing methods. We developed our training and test datasets using the miRTarBase repository, gene expression data from GEO and quantitative proteomics data based on pSILAC [146]. GEO is the largest source of microarray, RNA-seq and other forms of high-throughput genomics data [132]. We choose data from two microarray-based experiments - GSE6838 [141] and GSE8501 [143], and one pSILAC dataset [133]. As recommended in [144, 145], we remove the genes for which the expression levels are below the median in the control transfection experiments. We also remove ambiguous miRNA targets for which the absolute decimal logarithm of their gene expression ratio (transfected/control, fold change) is lower than 0.1. The targets are assumed to be functional when the $\log_{10}$ (fold change) is smaller than -0.1, and to be non-functional when the $\log_{10}$ (fold change) is larger than 0.1. The miRTarBase repository provides the largest number of curated

miRNA target samples which are validated by low-throughput assays, including both miRNA-mRNA interactions (the mRNA is repressed by a given miRNA) and miRNA-target duplexes (the miRNA binds a specific position on the target mRNA) [60]. We selected this repository because it has the largest number of functional (positive) miRNA-target duplexes and non-functional (negative) genes/sites that do not interact with a given miRNA. The duplex information is validated through gene mutation method which is regarded as strong evidence. Moreover, we only use the functional targets and non-functional genes for a given miRNA that are validated by at least one low-throughput assay, such as qPCR, luciferase assays and western blot. We also limit the data to human, which is consistent with the criteria to select the evaluated target predictors.

Table 5-1 Summary of the TRAINING and TEST datasets

Contents summarize criteria used to select the data from the sources. For the non-canonical targets in each dataset we list the number of miRNAs and genes and the corresponding number of functional (# fun) miRNA-mRNA pairs and non-functional (# nfun) pairs that were validated not to interact, separately for the 5mer and 6mer targets.

Dataset	Contents	Scope of annotation	6mer targets				5mer targets			
			# miRNAs	# genes	# fun	# nfun	# miRNAs	# genes	# fun	# nfun
TRAINING	miRTarBase in 2012 and earlier Microarray (random 10% from top 1%) pSILAC (random 10% from top 1%)	Gene/gene site Gene Protein	86	139	86	73	73	173	80	116
TEST_gene	miRTarBase_gene after 2012 Microarray (top 1% excluding TRAINING) pSILAC (top 1% excluding TRAINING)	Gene Gene Protein	42	109	69	58	51	177	74	148
TEST_expression	Microarray (top 50% excluding TRAINING) pSILAC (top 50% excluding TRAINING)	Gene Protein	27	3937	2481	2151	27	7923	4302	5728
TEST_duplex	miRTarBase_duplex after 2012	Gene site	15	13	11	5	5	6	2	4

We use experimentally validated targets and non-functional genes from miRTarBase that were published after 2012 to perform an unbiased empirical assessment. We utilize the older data to design our predictive models. The corresponding TRAINING dataset includes validated samples from 2012 or earlier from miRTarBase and 10% of randomly selected data from the top 1% of genes with the highest absolute $\log_{10}$ (expression fold change) values from the GEO and pSILAC datasets. We developed three TEST datasets using the same three data sources which are independent from the TRAINING dataset (they do not share the same sequences). TEST_gene and TEST_expression are utilized to assess the prediction of target mRNAs for a given miRNA.

TEST_expression is the biggest and includes 50% of genes with largest expression changes from GSE6838, GSE8501 and pSILAC test datasets at the transcriptome/proteome level, excluding the genes used in the TRAINING dataset. 50% is chosen to balance the size of the dataset and the reliability of the annotations. TEST_gene includes more reliable samples, which combine the top 1% of genes with highest absolute $\log_{10}$ (fold change) from the two microarray and one pSILAC test datasets (again, excluding the genes from TRAINING), and all validated miRNA-mRNA pairs and mRNA genes not repressed by a given miRNA (non-functional) that were deposited in miRTarBase after 2012. TEST_duplex is used to assess the prediction of miRNA binding sites on mRNAs and includes the miRNA-target duplexes from miRTarBase validated after 2012 and all non-functional sites that were validated as not bound by a given miRNA. We include all non-functional sites owing to their overall small count. Details are given in Table 5-1.

5.4 **ncMirTar Method**

ncMirTar performs predictions in two steps: (1) the input miRNA and 3' UTR of the mRNA sequences are converted into a set of numerical features; and (2) the features are inputted into one of two models designed for the prediction of the 6mer and 5mer targets. The motivation to use two models comes from Figure 5-1A that shows that about half of miRNA targets are non-canonical and among them 76% interact via 6mer and 5mer matches in their seeds. By screening the TRAINING dataset we found that fraction of native functional 6mer interactions among all 6mer interactions is much higher than for 5mer interactions. Furthermore, values of features computed for the 6mer- and 5mer-based interactions are also different. Consequently, we design different predictive models for each of the two types of the non-canonical targets. We note that current prediction methods are based on a single model, which means that they do not accommodate for the multimodality of the miRNA targets. Arguably, our design better captures the intrinsic properties of the two types of targets. Moreover, in our design we generate the input features by considering all types of inputs that were used in the previous predictors and we also introduce and use a new type of input. We use an empirical approach to maximize predictive performance for each of the two predictive models by combining these various features.

5.4.1 Features and feature selection

We consider six types of inputs and nine subtypes of inputs (some types are subdivided) to generate the features (Table 5-2). The six types include WC base pairing in the seed region and along the entire (predicted via alignment) miRNA-mRNA duplex, accessibility, conservation, nucleotide composition, position of the target sites, and occurrence of seeds or target sites. The first three input types are commonly used; 37 out of 38 methods summarized in the most recent review [111] employed at least one of these three types; four methods used all three types together: TargetScan [23], SVMicrO [75], mirSVR [144] and miRmap [145]. The position and the subtype of occurrence related to the target abundance on mRNA were also used in TargetScan. The nucleotide composition is often calculated for variety of predictive purposes. We also explored two novel subclasses of inputs: conservation of non-homologous genes and target ranking. The non-homologous gene conservation, which examines the occurrence of a gene fragment within the same genome, is parallel to the homologous gene conservation which looks at the occurrence of a gene fragment in the same gene across different species. Target ranking is a new idea based on an assertion that the miRNA-mRNA pairs for which seed types are at the top in ranking (i.e., have more WC pairs relative to other pairs with either different miRNAs or different mRNAs) are more likely to interact with each other (to be functional). In other words, we rank seed type associated with the current prediction among all putative miRNA-mRNA duplexes associated with the same miRNA and/or the same mRNA. We sort the seed types from 8mer, 7mer-8, 7mer-A1 (canonical) to 7mer-1, 6mer-2, 6mer-3, 6mer-1, 5mer-2, 5mer-3 and 5mer-4 (non-canonical). The nomenclature for the non-canonical seed type is *x*mer-*y*, where *x* is for the maximum count of consecutive WC base pairs in the seed and *y* is for the starting position of the matches from 5' end of miRNAs. We calculate three features in this group: (1) ranking of the seed type of the given miRNA_A and mRNA_B among the seed types of miRNA_A and mRNA_any, normalized by the count of mRNA transcripts; (2) ranking of seed type of the given miRNA_A and mRNA_B in the seed types of miRNA_any and mRNA_B pairs, normalized by the count of miRNAs; and (3) product of the two normalized rankings. Table 5-2 gives detailed description of all features.

Table 5-2 Description of the considered features

Type of input	Subtype of input	Description of features computed from a given type/subtype of input	Count	Total count
Complementary WC base pairing		Binding score in the seed from position 1-8	1*8=8	33
		% of WC and % of WC+GU in the first 4 nts in the seed, in the seed, in the non-seed, in position 13-16, in the whole region	2	
		% of WC, % of WC+GU, % of GU, maximum consecutive WC and maximum consecutive WC+GU in the seed, in the non-seed, in position 13-16, in the whole region	5*4=20	
		Normalized position of the first WC and last WC in the seed	2	
		Seed type	1	
Composition		Nucleotide composition in the seed, non-seed and whole region	4*3=12	54
		Nucleotide pairs composition in the seed, non-seed and whole region	14*3=42	
Accessibility	Minimum free energy (MFE)	Normalized MFE of the seed of miRNA-mRNA site duplex, entire miRNA-mRNA site duplex, mRNA site extended 10 nts upstream and downstream, mRNA site extended 30 nts upstream and downstream	5	7
		MFE difference between the duplex and mRNA site with 10 nts and 30 nts extension in both directions	2	
	AU content	AU content upstream, downstream and both ways from mRNA site	3	3
Conservation	Homologous conservation	% of conserved nucleotides, and conserved and paired nucleotides in the seed and non-seed regions	2*2=4	8
		Normalized score of conserved nucleotides, and conserved and paired nucleotides weighted by their positions in the seed and non-seed regions	2*2=4	
	Non-homologous conservation	% of conserved non-homologous genes and gene fragments	2	10
		Normalized maximum conserved nucleotides, and conserved and paired nucleotides in the seed and non-seed regions	2*2=4	
		Normalized sum of conserved nucleotides, and conserved and paired nucleotides in the seed and non-seed regions	2*2=4	
Target occurrence	Target abundance	Frequency of repeats of nucleotide 2-7 of the target seed in 3'UTR of the transcript, in the entire transcript, and in the whole genome	3	4
		# of predicted target sites in the mRNA	1	
	Target ranking	Ranking of the seed type of the predicted miRNA-mRNA pairs in all pairs associated with this miRNA	1	3
		Ranking of the seed type of the predicted miRNA-mRNA pairs in all pairs associated with this mRNA	1	
		Product of the above two rankings	1	
Target site position		Normalized position of the predicted target site in the 3'UTR of the transcript	1	1

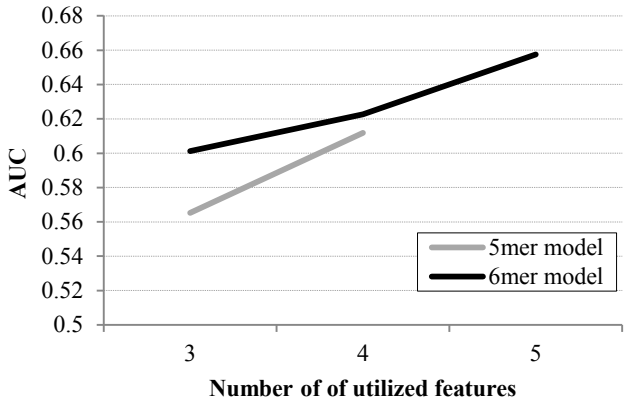


Figure 5-2 Predictive quality of 6mer and 5mer models with increase in the number of included features that were considered during feature selection. The results were computed based on three-fold cross validation on the TRAINING dataset.

Given the large number of 123 considered features and the fact that they may not be predictive and could be redundant (e.g., features generated in the same subclass are likely to be correlated with each other), we perform an empirical, two-step feature selection to find a well-performing subset of the features. In the first step, we remove the features which have lower predictive quality. The features are sorted by their average (over three-fold cross validation) *AUC* and correlation with the native annotations using the TRAINING dataset. The features with *AUCs* or correlation coefficients that are lower than the median values over all features are removed. In the second step, we take advantage of the interdependence between features and reduce redundancy by choosing at most one feature from each subtype of input and removing the remaining features in that subtype; some subtypes may have no features left after the first step. In this step, the features sorted by the *AUC* values are grouped into the subtypes and we empirically search for an interdependent feature set that gives highest predictive performance when applied together using SVM classifier based on three-fold cross validation on the TRAINING dataset. We start with a set of three features, selected as the best triplet among all possible combinations. Next, we add one additional feature (from the remaining input subtypes) if that results in improved predictive performance by at least 0.02 of *AUC*. The trends that show how *AUC* values change with the increase in the number of features for the 6mer and 5mer models are plotted in Figure 5-2. The final 6mer and 5mer models outperform the initial models with three features in *AUC* values by 9.4% and 8.3%, respectively. We use the cross validation and limit our selection

to one feature from each subclass to assure that the resulting feature set is small, which reduces likelihood of over-fitting the TRAINING dataset.

Table 5-3 Summary of the considered and selected types of inputs and the corresponding features

Type of input	Subtype of input	Number of features				
			6 mer		5 mer	
		All considered	After step 1 low quality removed	After step 2 redundant removed	After step 1 low quality removed	After step 2 redundant removed
Nucleotide composition		54	15	1	12	1
Base pairing		33	15	1	10	1
Accessibility	Free energy	7	5	0	3	0
	AU content	3	3	1	3	1
Conservation	Homologous	8	0	0	6	1
	Non-homologous	10	2	0	7	0
Target occurrence	Target abundance	4	3	0	0	0
	Target ranking	3	3	1	0	0
Target site position		1	1	1	1	0

Following the two-step feature selection procedure, 47 and 42 features pass the first step, and 5 and 4 features from 6 subtypes of inputs are selected in the second step for the 6mer and 5mer model, respectively (**Error! Reference source not found.**). The novel feature based on the target ranking was selected. The other selected features are commonly used in prediction of canonical targets and they include complementarity WC base pairing, accessibility and conservation. Based on the recent review [111], the base pairing was used by 37 out of 38 existing methods, 31 methods employed one or both subtypes of the accessibility, and half of the predictors utilized conservation. Importantly, each of the two models in ncMirTar uses a different set of features which agrees with our observation that predictions for 6mer and 5mer should be individualized. Moreover, we use an empirical approach to maximize the predictive performance on the non-canonical targets by selecting only the predictive and non-redundant subset of new and previously used inputs.

5.4.2 Architecture of the ncMirTar method

A detailed outline of ncMirTar is shown in Figure 5-3. This is a sequence-based method, which means that it only requires miRNA and mRNA sequences as inputs. We also offer a pre-computed database of ncMirTar's predictions which is based on miRNAs collected from the Release 21 of miRBase [59] and mRNAs from Release Nov. 2014 of RefSeq [187]. First,

ncMirTar slides the "seed" of the input miRNA over the 3' UTR of mRNA transcript to fold miRNA and mRNA segments into duplexes. The position of 3' UTR is extracted from Genbank using BioPerl toolkit [188]. The duplex structure is estimated using a revised Smith-Waterman algorithm, which is a dynamic programming algorithm to perform local sequence alignment [189]. The revised algorithm rewards the WC base pairs and GU wobbles instead of the same residue matches for the alignment, and retains the penalty for mismatches and gaps. For each pair of miRNA and mRNA, we process the interaction site with the maximum number of WC base pairs in the seed. This means that if we find a canonical site then this mRNA is added into a list of targets for the prediction with TargetScan; we note that our procedure to find canonical targets is the same with TargetScan. If a 6mer site is found with no canonical sites then 5mer sites will not be searched for and ncMirTar's 6mer model will be used. If no canonical and 6mer sites are found but 5mer site(s) is found then we use the ncMirTar's 5mer model. We assume that miRNA does not interact with the mRNA if there are no canonical, 6mer and 5mer sites. We use different models to predict 6mer and 5mer sites. Both models utilize SVM and a different set of features. Based on the recent review [111], besides SVMs Bayesian statistical models and artificial neural networks are also commonly used as predictive model for finding miRNA targets . Bayesian modeling is a statistics-based method and requires a large number of training samples compared to the number of features in order to achieve good predictive performance. Neural networks may get stuck in local minima/maxima when being trained. We employed SVM which does not require a large number of training samples and finds a global optimum. We computed SVM with the LIBSVM package using the default Gaussian kernel and parameters c = 200 [190]. Each of the two SVM models outputs their own scores (propensities) and we convert these scores into rankings on the whole genome; the scores range from -100 to 100 with positive values for putative functional targets and negative for non-functional genes for a given miRNA. We also binarize the ranked scores using threshold of 30, which gives the highest *AUC* value on the TRAINING dataset. Lastly, we note that the entire design of ncMirTar was done using the TRAINING dataset, which is independent (does not share the same sequences) from the TEST datasets that we used for the comparative analysis.

Our first-of-its-kind method for prediction of the non-canonical miRNA target implements two novel ideas: (1) predictions are based on two models that are specialized for two prevalent non-canonical seed types; and (2) the inputs consist of a small empirically selected set of features that

include a new feature type. Next, we empirically demonstrate that these two ideas significantly contribute to the predictive performance of ncMirTar.

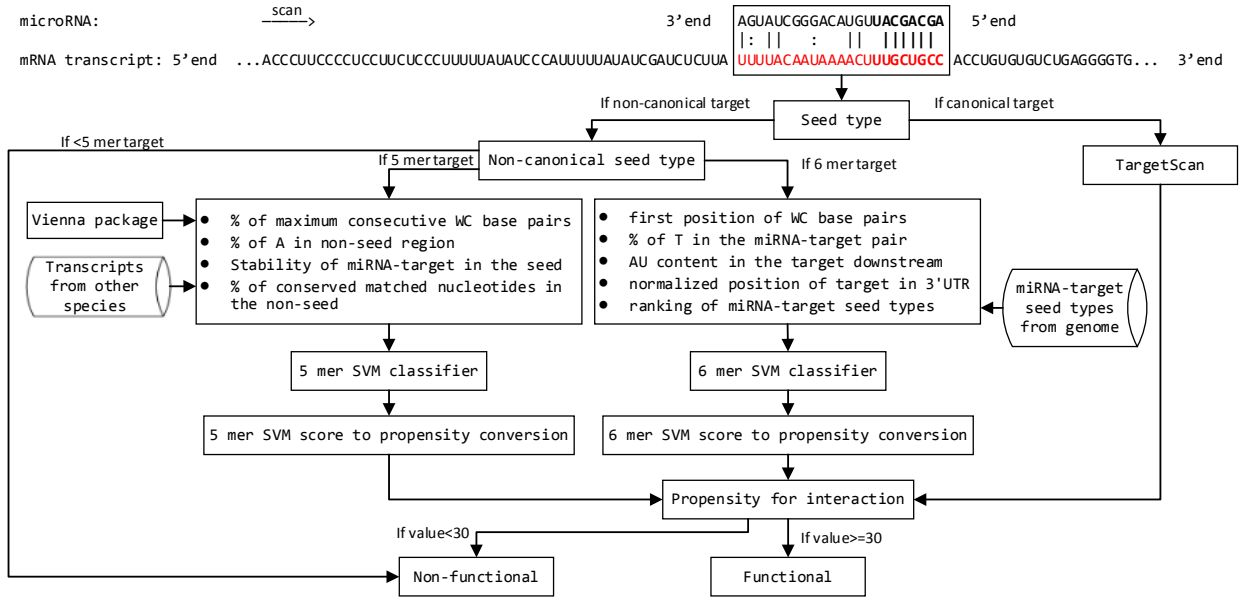


Figure 5-3 Architecture of the ncMirTar predictor

5.4.3 Assessment of novel aspects of ncMirTar

Figure 5-1A shows that the proportions of different non-canonical seed types are consistent across the five datasets and that on average over 76% of the non-canonical targets are 6mers and 5mers. We do not consider the remaining 24% of the non-canonical targets since the number of WC pairs in their seed region is too low to allow for an accurate prediction. The values of the features computed for the 6mer- and 5mer-based miRNA-mRNA interactions are different and thus we designed two models for these two non-canonical types of targets. We compare the ncMirTar's design which is based on the two models with the design based on a single model that combines both seed types that was built using the same design procedure (Figure 5-4A). The ncMirTar (dark grey bar) improves over the single model design (red bar) by 3.9% in *AUC* and this difference is statistically significant (p-value<0.01).

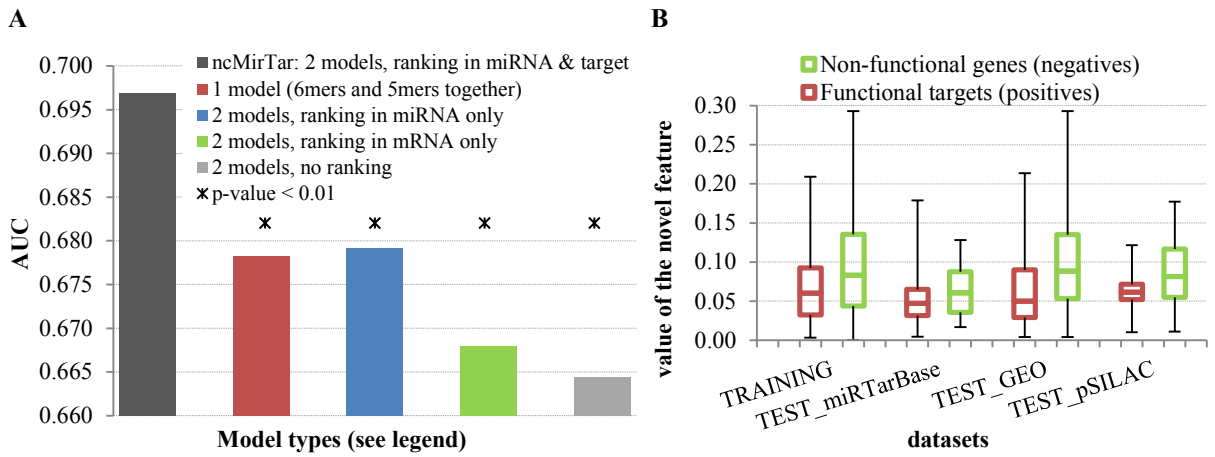


Figure 5-4 Analysis of the impact of the two-model design and inclusion of the new feature type
(a) Comparison of *AUC* values obtained by using the ncMirTar (dark gray bar), which uses two models for 5mer and 6mer targets and novel feature based on ranking in miRNA and mRNA; design that uses a single model and ranking in miRNA and mRNA (red bar); and three designs with two models and ranking only in miRNA (blue bar), only in mRNA (green bar), and with no novel feature based on the ranking (light gray bar). Stars at the top of the bars indicate that difference in *AUC* values between ncMirTar and other types of models is significant at *p*-value < 0.01. (b) Values of the novel feature based on ranking of miRNA-target seed types for the functional targets (red box-plots) and non-functional targets (green box-plots) on the four datasets. The boxes give the first quartile, median and the third quartile of the values; the two whiskers show the minimum and maximum values.

In the design of ncMirTar we explored new types of features and combined them with features that were inspired by inputs used in the other methods. The final set of input features was empirically selected from both prior and novel features. The novel feature used in ncMirTar is based on ranking of miRNA-target seed types, which is a product of ranking associated with the given miRNA and with the given target (see section 5.4.1). We compare the predictive performance of ncMirTar (dark gray bar in Figure 5-4A) with the design that does not use the novel feature (light gray bar), and which uses the novel feature based on ranking only in either mRNA (green bar) or miRNA (blue bar). The *AUC* value of ncMirTar is significantly larger (*p*-value<0.01) than the *AUC* when the new feature was removed or modified. We compare values of this new feature between functional and non-functional miRNA-mRNA samples in the four datasets (Figure 5-4B). The median values for the functional targets are consistently lower than those for the non-functional genes, i.e., the ranking of miRNA-target seed types in the functional targets is generally higher than the ranking in the non-functional genes, which agrees with intuition. The average (over the four datasets) of the median values for the functional pairs is 0.055, which corresponds to the ranking in the top $\sqrt{0.055}$=23.4%. Given that on average close to 50% of miRNA targets are canonical (Figure 5-1a), our feature relies on the observation that

the considered non-canonical interaction is more likely to be functional when the given miRNA or target gene has fewer targets or miRNAs that form canonical pairs (ranking of the non-canonical interaction is sufficiently low).

5.5 Comparative evaluation of ncMirTar

We empirically compare ncMirTar with several representative sequence-based predictors. The requirements to filter out existing methods are similar and a bit more liberal compared with those used in Table 4-6. The selected methods have to be conveniently accessible to the end users via a webserver or a pre-computed database. They also have to cover human, predict target binding sites, and generate propensities (real-values that quantify likelihood of the interaction). Using these filters we selected nine methods (see Table 5-4). We use the latest versions of these methods, except for PicTar2 which is substantially different from PicTar and no longer qualifies as a sequence-based predictor. PicTar v2005 was first published in 2005; five methods including TargetScan 6.2, miRanda 2010, EIMMo3, MultiMiTar and mirTarget2 v4 were proposed or updated between 2010 and 2012; two in 2013: DIANA-microT-CDS and miRmap v1.1, and MIRZA-G in 2015. These nine methods use a diverse set of predictive models, with four that utilize heuristic scoring functions and five that use the machine learning models including Bayesian classifier, SVM and regression. DIANA-microR-CDS, miRmap and MIRZA-G were built based on gene expression data, while the other methods were derived based on the low-throughput experimentally validated data.

Table 5-4 Summary of the criteria used to select methods for the empirical assessment

The 'covered species' include chicken (c), drosophila (d), dog (g), human (h), mouse (m), nematode (n), rat (r) and zebra fish (z). A selected method must at minimum predict for human and mouse. The 'outputs' summarize the format and scope of the outputs generated by a given method. The formats of outputs include real-valued propensity of the miRNA-mRNA interactions (probability or score) and binary outcome (binding vs. non-binding). A selected method must provide the more informative real-valued propensity. The target sites can be tracked by the predicted position on the mRNA sequence, by the matching of the seed on the mRNA sequence, by both of these options, or the output does not allow the tracking. We did not consider methods that do not allow the tracking since this is required to perform analysis at the miRNA-mRNA duplex level. "Ability to predict non-canonical targets" divides the methods into three groups. "×" denotes methods cannot predict non-canonical targets. "o" represents methods predict non-canonical targets but did not evaluate their predictions. "✓" is for methods that evaluate predictions of non-canonical targets but were not designed for them. The 'notes' list extra features or important drawback of the predictors. Eight out of nine evaluated predictors are capable of batch predictions, which facilitated our tests. miREE did not predict any targets in either TEST_duplex or TEST_gene datasets and thus was excluded. Although MIRZA-G cannot track the target binding site, this is the latest predictor which is also capable of predicting the non-canonical miRNA targets. Thus, we made an exception and included this method; we used its gene prediction for the duplex prediction on the TEST_duplex dataset. The features that motivated exclusion of a given method from the empirical evaluation are given in italics on a gray background. Names of the selected methods are given in bold font.

Predictor	Covered species	Outputs		Ability to predict non-canonical targets	Notes
		Score	Target site tracking		
TargetScan v6.2	c, d, h, m, z	probability	both	×	batch search
RNAhybrid v2.1	any	*binary*	sequence	o	*always predicts a target*
DIANA-microT-CDS	d, h, m, n, r	probability	both	o	batch search
miRanda v2010	d, h, m, n, r	probability	sequence	✓	batch search
PicTar v2005	d, h, m, n	probability	seed sequence	o	
MicroInspector v1.5	any	*binary*	both	o	
RNA22 v2.0	d, h, m, n	*binary*	both	o	
PITA v2007	d, h, m, n	*binary*	position	×	
STarMir v2007	h, m	probability	position	o	*long runtime*
EIMMo3	d, h, m, n, r	probability	position	o	batch search
TargetRank v2007	h, m	probability	*none*	o	
MirTarget2 v4.0	c, g, h, m, r	probability	both	×	batch search
TargetMiner v2012	h	*binary*	seed sequence	o	
TargetSpy v1.0	c, d, h, m, r	*binary*	sequence	×	
RepTar v1.2	h, m	*binary*	sequence	o	
MultiMiTar	h	score	sequence	o	batch search
miREE	c, h, m, n, r, z	probability	both	o	*did not predict targets*
miRcode v11	h	*none*	position	×	
miRmap v1.1	c, e, h, m, o, r, w, z	probability	both	o	batch search
MIRZA	any	probability	sequence	✓	*length of input gene sequences limited to between 30 and 50nt*
MIRZA-G	h	probability	*none*	✓	batch search

We collected predictions for these methods using either their online webservers or downloadable pre-computed predictions. We recorded their predicted binding targets (sequences or positions) and the corresponding propensities. Following ref. [111], a given pair of miRNA and mRNA from the TEST_gene and TEST_expression datasets was predicted as a functional interaction if

the mRNA was predicted as a target using the corresponding miRNA; the propensity was computed as the sum of scores generated by the predictive model for all predicted target sites for the given miRNA-mRNA pair; the remaining pairs were assumed to be non-functional and the corresponding propensity was set to 0. A given pair of miRNA and mRNA site in the TEST_duplex dataset was predicted as a functional interaction if the mRNA was predicted as a target using the corresponding miRNA and the difference between the predicted and actual position of the binding site was smaller than four nucleotides; the remaining pairs were predicted as non-functional and the corresponding propensity was set to 0.

We assess the predictive performance of these methods on three benchmark datasets that include data collected from different sources: low-throughput and high-throughput experiments from miRTarBase, and from GEO and pSILAC. Only targets validated by low-throughput after 2012 are used for test, since six out of nine methods that we compare with were published in 2012 or earlier and the other three did not use data from the curated databases for training. These datasets also focus on different levels of annotation. TEST_gene and TEST_duplex datasets are utilized for the assessment at the gene/protein level (to predict whether a given mRNAs interacts with a given miRNA) and the duplex level (to predict whether a given fragment on mRNA interacts with a given miRNA). TEST_expression dataset is used to assess the gene level prediction but at the transcriptome/proteome scale (to predict all possible mRNAs that interact with a given miRNA). Moreover, we analyze predictive performance on the complete set of all targets and separately for the non-canonical targets. We perform total of six evaluations considering non-canonical vs all targets for the TEST_ gene, TEST_duplex and TEST_expression datasets.

Following recent review from ref. [111], we evaluate the predictive quality using a comprehensive set of five measures. Four measures are used to assess binary predictions (functional/interacting vs. non-functional target): sensitivity (true positive rate (*TPR*)) and specificity that quantify fraction of correctly predicted functional and non-functional mRNA genes, Matthews correlation coefficient (*MCC*) that gives the overall predictive performance, and predicted-to-native positive rate (*PNR*) that quantifies the amount of predicted targets. The predicted propensities (real-valued scores that quantify likelihood of the miRNA-target interaction) are assessed using area under *ROC* curve (*AUC*) and we evaluate statistical significance of the differences in *AUC* between ncMirTar and the other considered methods. We

also investigate ability to predict highly repressed target genes based on the expression fold change of top 20% (with highest propensities) of predictions (*Expression_20%*). Except for *PNR*, the other measures were previously used to assess published predictors, although never altogether. Sensitivity and specificity were utilized to evaluate EIMMo [162], PACMIT [74] and miREE [167]; *MCC* was the main criterion to evaluate TargetMiner [78] and MultiMiTar [166]; many predictors use *ROCs*, such as MirTarget2 [147] and miRanda-mirSVR [144]. miRanda-mirSVR [144] and MIRZA-G [186] also evaluated expression level changes of their top predictions.

Comparison of predictions at the gene-level for the non-canonical targets (TEST_gene dataset) shows that ncMirTar offers the highest sensitivity; it correctly predicts the largest fraction of over 40% of the functional targets compared to 24.5% for the second best miRmap, and the average of 6.7% over the nine other methods (Table 5-5). At the low false positive rate (*FPR*) = 0.05 (specificity=0.95) ncMirTar offers high true positive rate (sensitivity) of 0.24, compared to 0.14 for the second best miRmap. Our predictor also secures the highest *MCC* value of 0.34 and improves over the second best miRmap by 0.13 (relative improvement = 100%*(0.34-0.21)/0.21=62%) (Figure 5-5A). Moreover, ncMirTar secures the highest *AUC* value (used to assess the propensities) of 0.705 which is significantly higher (p-value<0.01) than *AUCs* of the other nine methods, with the relative improvement of 22.0% compared to the runner-up miRmap (*AUC* = 0.578) (Figure 5-5A). Their *ROC* curves are given in Figure 5-5B. The *ROC* of ncMirTar is consistent above the other curves with the entire *FPR* for a large margin. The *PNR* value quantifies the total number of predicted functional targets and reveals that ncMirTar predicts the most at 55.8% (Figure 5-5A). The other methods under-predict by a large margin, where TargetScan and MirTarget2 predict no non-canonical targets. Only three methods achieve *PNR*>20% on this dataset. We examine their ability to predict highly repressed target genes among their predicted functional genes using *Expression_20%* (gene expression fold changes of the top 20% of predictions). *Expression_20%* = -0.53 for ncMirTar, which means that the top 20% of predicted genes are repressed on average by 70.5% ($1-10^{-0.53}=0.705$). The top 20% of the genes predicted with DIANA-microT-CDS and miRmap are repressed by 59.4% and 53.7%, respectively (Table 5-5). This suggests that propensities provided by ncMirTar for the non-canonical targets allow selecting a set of more repressed genes by the corresponding miRNAs.

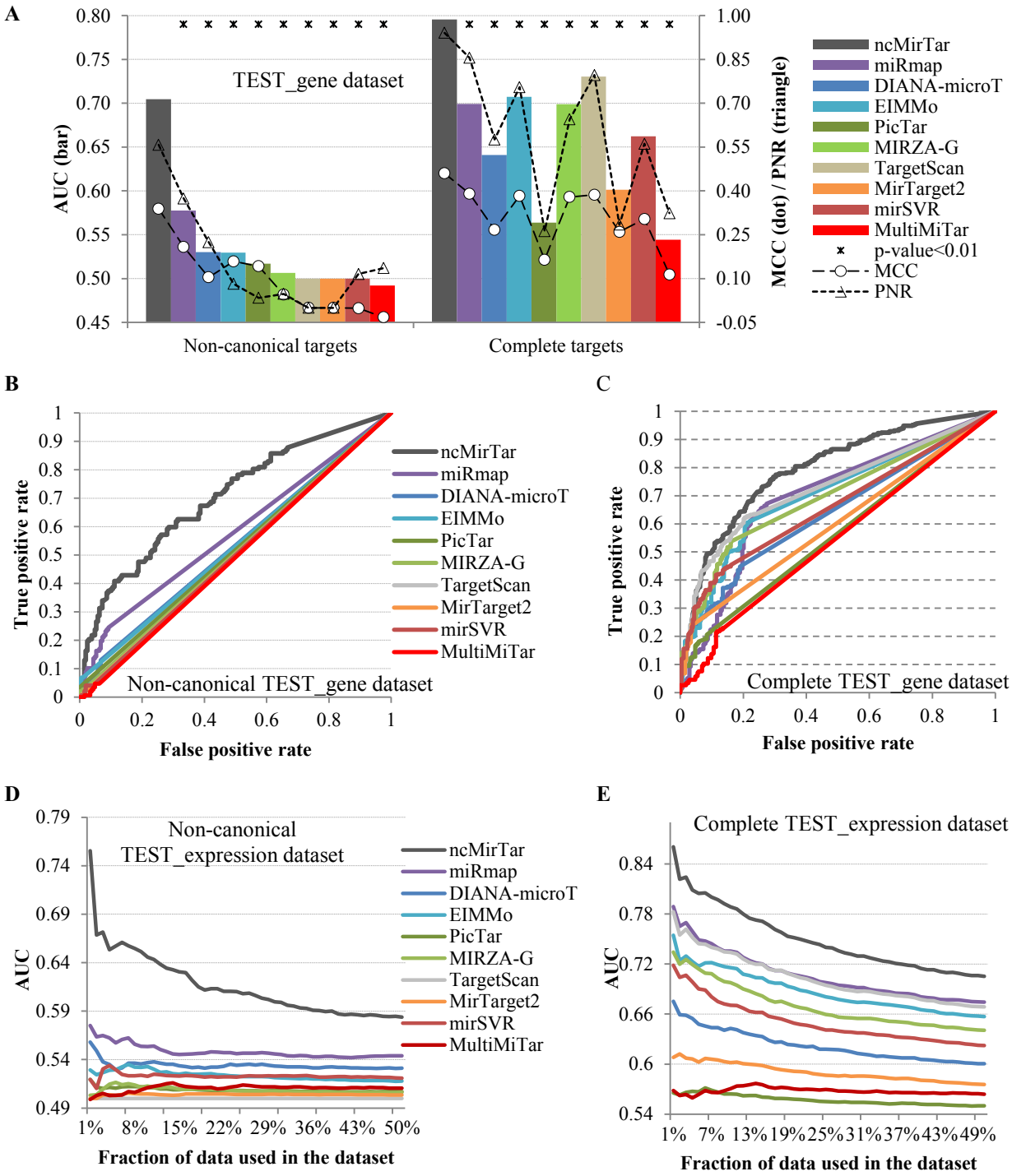
Table 5-5 Comparative evaluation of ncMirTar and other representative predictors at the gene level (TEST_gene dataset) and at the duplex level (TEST_duplex dataset)

We measure area under the ROC curve (*AUC*), Matthews Correlation Coefficient (*MCC*), sensitivity (*Sen.*), specifity (*Spe.*), predicted-to-native functional target ratio (*PNR*), expression fold changes of the top 20% predictions ($Expression_{20}$) and statistical significance of differences in *AUC* values between ncMirTar and other methods. Methods are sorted in the descending order by their *AUC* values for non-canonical target prediction on TEST_gene. The best value of each measure across all predictors is shown in bold font.

Prediction type	Predictor	*AUC*	*p*-value	*MCC*	Sen.	Spe.	PNR	$Expression_{20}$
Non-canonical targets at the gene level (TEST_gene dataset)	ncMirTar	**0.705**		**0.339**	**0.401**	0.889	**0.558**	**-0.530**
	miRmap	0.578	6.7E-06	0.208	0.245	0.908	0.374	-0.334
	DIANA-microT-CDS	0.530	3.3E-08	0.104	0.129	0.932	0.224	-0.391
	EIMMo	0.529	2.2E-07	0.159	0.068	0.990	0.082	
	PicTar	0.517	1.1E-07	0.142	0.034	**1.000**	0.034	
	MIRZA-G	0.506	1.2E-07	0.045	0.027	0.986	0.048	
	TargetScan	0.500	2.8E-08	0.000	0.000	**1.000**	0.000	
	MirTarget2	0.500	2.8E-08	0.000	0.000	**1.000**	0.000	
	miRanda-mirSVR	0.500	5.8E-09	-0.002	0.048	0.952	0.116	
	MultiMiTar	0.492	1.1E-08	-0.032	0.048	0.937	0.136	
All targets at the gene level (TEST_gene dataset)	ncMirTar + TargetScan	**0.795**		**0.460**	**0.750**	0.715	**0.941**	-0.606
	miRmap	0.699	1.2E-08	0.390	0.672	0.726	0.855	**-1.076**
	DIANA-microT-CDS	0.641	2.1E-10	0.267	0.449	0.810	0.576	-0.936
	EIMMo	0.707	8.4E-08	0.383	0.608	0.781	0.755	-0.966
	PicTar	0.564	1.5E-13	0.164	0.206	0.916	0.262	-0.112
	MIRZA-G	0.699	1.1E-08	0.379	0.537	0.839	0.645	-0.799
	TargetScan	0.731	1.3E-07	0.387	0.635	0.759	0.797	-0.606
	MirTarget2	0.601	2.3E-12	0.260	0.250	**0.949**	0.284	-0.808
	miRanda-mirSVR	0.662	1.3E-08	0.304	0.453	0.839	0.561	-0.667
	MultiMiTar	0.544	4.5E-13	0.114	0.230	0.861	0.324	-0.655
Non-canonical targets at the duplex level (TEST_duplex dataset)	ncMirTar	**0.829**		**0.437**	**0.538**	0.889	**0.615**	
	miRmap	0.577	4.8E-05	0.263	0.154	**1.000**	0.154	
	DIANA-microT-CDS	0.615	1.7E-04	0.331	0.231	**1.000**	0.231	
	EIMMo	0.538	1.5E-06	0.182	0.077	**1.000**	0.077	
	PicTar	0.500	1.5E-06	0.000	0.000	**1.000**	0.000	
	MIRZA-G	0.538	5.2E-06	0.182	0.077	**1.000**	0.077	
	TargetScan	0.500	1.5E-06	0.000	0.000	**1.000**	0.000	
	MirTarget2	0.500	1.5E-06	0.000	0.000	**1.000**	0.000	
	miRanda-mirSVR	0.577	8.9E-06	0.263	0.154	**1.000**	0.154	
	MultiMiTar	0.577	3.7E-05	0.263	0.154	**1.000**	0.154	
All targets at the duplex level (TEST_duplex dataset)	ncMirTar + TargetScan	**0.699**		0.270	**0.927**	0.296	1.081	
	miRmap	0.648	3.2E-03	0.140	0.724	0.444	0.846	
	DIANA-microT-CDS	0.679	3.0E-01	**0.279**	0.585	0.778	0.634	
	EIMMo	0.544	1.1E-05	0.110	0.756	0.370	0.894	
	PicTar	0.526	1.3E-06	0.054	0.285	0.778	0.333	
	MIRZA-G	0.628	3.0E-05	0.186	0.740	0.481	0.854	
	TargetScan	0.672	4.1E-04	0.210	0.870	0.333	**1.016**	
	MirTarget2	0.538	1.2E-06	0.086	0.325	0.778	0.374	
	miRanda-mirSVR	0.562	1.0E-06	0.078	0.472	0.630	0.553	
	MultiMiTar	0.552	2.3E-05	0.067	0.260	**0.815**	0.301	

We also assess predictive performance at the gene-level on the complete TEST_gene dataset that includes both canonical and non-canonical targets. We combine the ncMirTar's predictions for

the non-canonical targets with the TargetScan's predictions for the canonical targets; TargetScan offers the best predictive performance measured with *AUC* among the nine other methods on the TEST_gene dataset, which is consistent with the results in a recent review [111]. To be noticed, ncMirTar does not predict any canonical targets and TargetScan does not predict any non-canonical targets, so the predictions of the two methods would not overlap. The two measures of the overall predictive performance, *AUC* and *MCC*, indicate that combination of ncMirTar and TargetScan outperforms the other methods (Figure 5-5A and Table 5-5). The corresponding *AUC* value = 0.795 compared to 0.731 for the best solo TargetScan, and *MCC* value = 0.460 compared to the second best miRmap at 0.390. The *AUC* of the ncMirTar with TargetScan combo is also significantly higher (p-value<0.01) than the *AUCs* of the other methods (Figure 5-5A). This demonstrates that adding predictions for the non-canonical targets generated by ncMirTar to the best performing methods for the prediction of canonical targets results in the substantial increase in the predictive performance. Their *ROC* curves are plotted in Figure 5-5C. ncMirTar coupled with TargetScan also achieves the highest sensitivity of 0.75. The runner-up miRmap and solo TargetScan also provide good sensitivity values at 0.67 and 0.64, respectively. PicTar and MirTarget2 are the two most conservative methods based on their highest specificity values; they predict only a few functional targets (sensitivity values equal 0.21 and 0.25, respectively) but with high success rate. The *PNR* values reveal that number of functional targets predicted by ncMirTar and TargetScan combo is the closest to the number of native targets (94.1%), with miRmap being the second best (85.5%). *PNR* of TargetScan and EIMMo equals 79.7% and 75.5%, respectively, while the other methods under-predict the functional genes by a large margin. The *Expression_20%* values of the ncMirTar and TargetScan combination are worse on the complete dataset compared to some other methods, which stems from the relatively poor performance of the solo TargetScan. Consequently, the best option to find highly repressed non-canonical genes is ncMirTar and for the canonical genes are miRmap, EIMMo and DIANA-microT-CDS (Table 5-5).


A
TEST_gene dataset
AUC (bar)
MCC (dot) / PNR (triangle)
Non-canonical targets
Complete targets
ncMirTar
miRmap
DIANA-microT
EIMMo
PicTar
MIRZA-G
TargetScan
MirTarget2
mirSVR
MultiMiTar
p-value<0.01
MCC
PNR
B
True positive rate
False positive rate
Non-canonical TEST_gene dataset
C
Complete TEST_gene dataset
D
Non-canonical
TEST_expression dataset
AUC
Fraction of data used in the dataset
E
Complete TEST_expression dataset

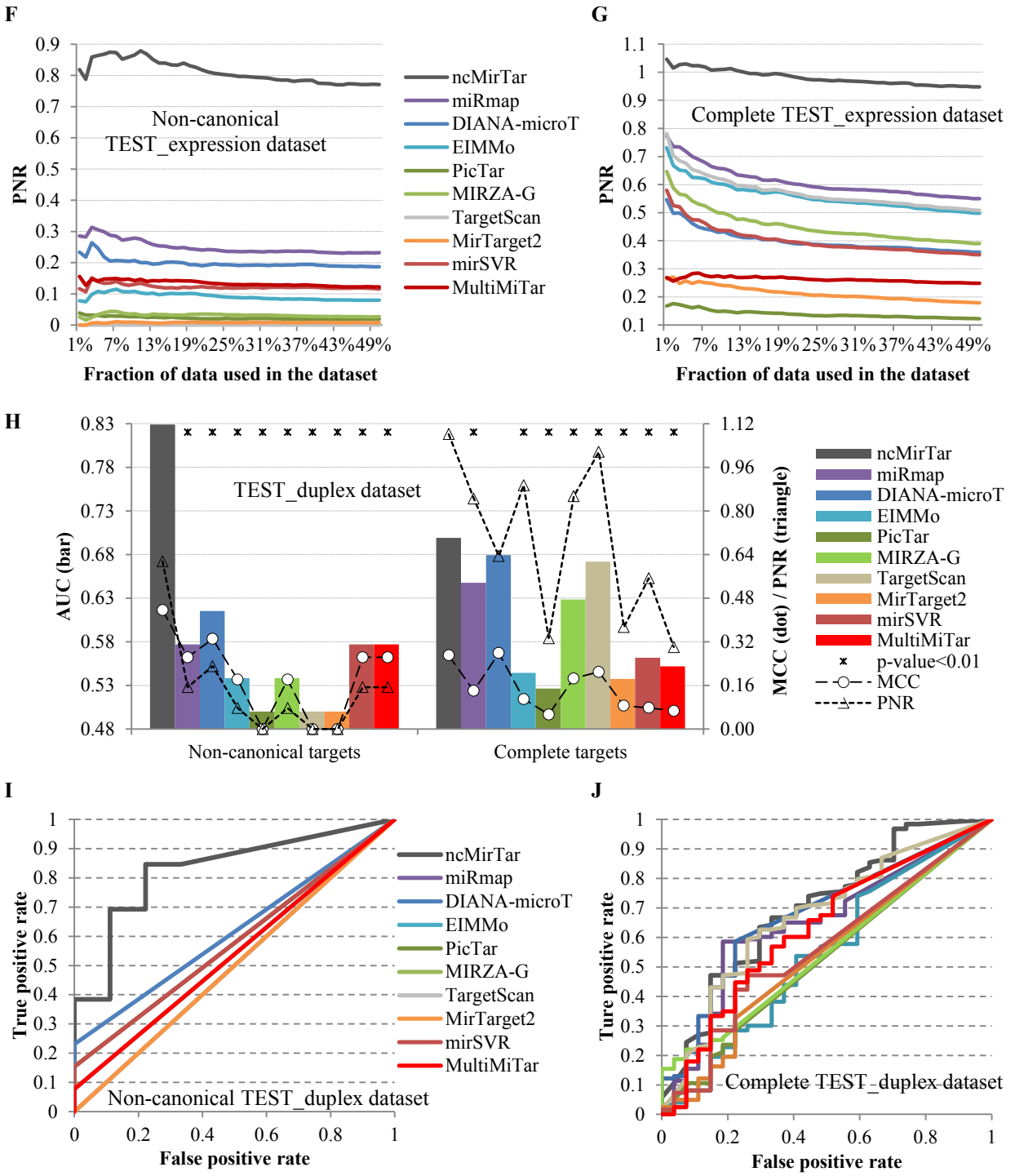


Figure 5-5 Comparative evaluation of ncMirTar and other representative predictors on the TEST_gene and TEST_expression datasets

Methods are sorted according to their *AUC* values on the non-canonical TEST_gene dataset. (A) and (H) summarize the *AUC* (bars), *MCC* (dashed line with dot markers) and *PNR* (dotted line with triangle markers) values on the non-canonical targets and complete (both canonical and non-canonical) TEST_gene and TEST_duplex datasets, respectively. Stars at the top of the figure indicate that difference in *AUC* values between ncMirTar and a given methods is significant at *p*-value < 0.01. (B), (C), (I) and (J) plot *ROCs* of the ncMirTar and nine other popular predictors on the TEST_gene and TEST_duplex datasets for non-canonical predictions and complete predictions. (D) and (E) show the *AUC* values in function of the fraction of data (*x*-axis) used to build the TEST_expression

dataset for the non-canonical targets and the complete TEST_expression dataset, respectively. (F) and (G) plot *PNR* values and the thresholds used to define the functional (most suppressed) and nonfunctional (most overexpressed) genes for the predictions on the TEST_expression dataset for non-canonical targets and complete targets.

Next, we evaluate predictions at the transcriptome/proteome scale on the TEST_expression dataset. Here, each miRNA is associated with a large number of native functional and non-functional genes/proteins together with their expression level changes which quantify confidence of annotations. We vary the amount of functional and non-functional genes based on a threshold used to include *x*% of data with largest changes in the expression levels; we consider the values of *x*% between 1% and 50% corresponding with the confidence levels from high to low. The repressed (with negative fold changes) and over-expressed (with positive fold changes) genes are annotated as functional and non-functional, respectively. *AUC* values at different thresholds *x*% are given in Figure 5-5D and Figure 5-5E. As expected, they decrease when genes with more ambiguous annotations (larger *x*% values) are included. *AUCs* are higher on the complete TEST_expression dataset (Figure 5-5E) compared to the results on the non-canonical targets (Figure 5-5D). ncMirTar always secures the highest *AUCs*, irrespective of the threshold value, when predicting the non-canonical targets. The improvements offered by ncMirTar are larger when evaluating on lower *x*% of data which has more reliable annotations (Figure 5-5D). The other methods cannot predict the non-canonical targets accurately given that their *AUCs* are < 0.58. Similarly, combination of ncMirTar and TargetsScan which is used to predict the complete dataset also consistently outperforms the other methods (Figure 5-5E). The predictions from the solo TargetScan and miRmap are also relatively good, with the *AUC* values ranging between 0.67 and 0.79. We also plot the *PNR* curves in Figure 5-5F and G. The number of predictions from ncMirTar or ncMirTar_TargetScan combo is most close to the number of actual targets. All the other method under-predict especially for the non-canonical predictions.

Finally, we assess the predictive performance at the duplex level on the TEST_duplex dataset (Figure 5-5H). Consistent with the results on the other two datasets, ncMirTar offers improved predictive performance for the prediction of the binding sites on the non-canonical targets when contrasted with the other nine methods. ncMirTar secures the highest values of *AUC* = 0.829 and *MCC* = 0.437, and improves over the second best DIANA-microT-CDS by 0.214 and 0.106, respectively (relative improvements of 34.7% and 32.1%, respectively). The improvements in the *AUC* values are statistically significant (p-value<0.01), which is also obvious from their *ROC*

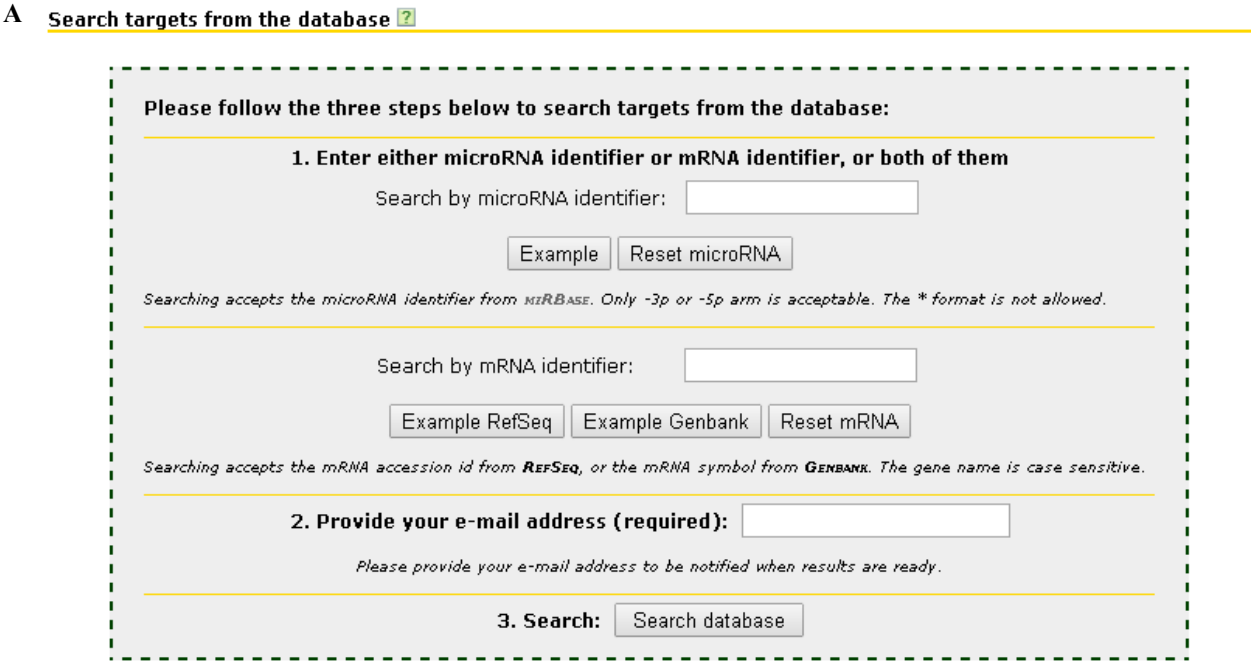
curves in Figure 5-5I. ncMirTar also secures the highest sensitivity (true positive rates) of 0.54, compared to 0.23 by the runner-up DIANA-microT-CDS. Although other methods have specificity = 1 (no false positives) on the non-canonical targets, their sensitivity values are very low with average of 0.09. In other words, ncMirTar trades a few false positive predictions for the substantial gain in sensitivity. Moreover, ncMirTar predicts the largest number of functional targets with *PNR* = 61.5%, although still under-predicting when compared with the number of native duplexes. However, the other methods under-predict by a much larger margin – their *PNR* values are no larger than 23.1%. The results on the complete TEST_duplex dataset that includes canonical targets reveal that combining ncMirTar with TargetScan again leads to improved predictive performance. The improvements are smaller compared to the other datasets since the number of non-canonical targets in TEST_duplex is lower compared to the canonical targets. However, the ncMirTar+TargetScan combo still secures the highest sensitivity = 0.927 and *AUC* = 0.699, and improves over the second best solo TargetScan by 0.057 (relative improvement of 6.5%) in sensitivity and over DIANA-microT-CDS by 0.020 (relative improvement of 2.9%) in *AUC*. The increase in *AUC* between ncMirTar and the other methods is statistically significant (p-value<0.01), except for DIANA-microT-CDS (Figure 5-5H). Their *ROC* curves are provided in the Figure 5-5J. Additionally, the *PNR* of TargetScan that equals 102% is the closest to the perfect 100% while *PNR* for the combination of ncMirTar and TargetScan is also very good at 108%. The other methods under-predict the functional targets by between 10.6% (EIMMo) and 70% (MultiMiTar).

To summarize, we show that ncMirTar provides high quality predictions at the gene, transcriptome and duplex levels that are competitive when compared to the representative predictors in this field.

5.6 **Availability of ncMirTar**

The ncMirTar method is freely available at http://biomine-ws.ece.ualberta.ca/ncMirTar/. We provide access to a pre-computed and fast-to-query database of putative miRNA-mRNA interactions in the human and mouse genomes (Figure 5-6A). This database can be searched by miRNA ID based on miRBase nomenclature (query returns genome-wide predictions), by gene ID using either RefSeq's or GeneBank's nomenclature (query returns all interacting miRNAs), or for a specific miRNA-mRNA pair. Results of a given query include detailed predictions using

the ncMirTar's model for all non-canonical targets. For each considered miRNA these results include location of putative interaction sites, the corresponding putative miRNA-mRNA duplexes and their seed types, propensity score for each predicted duplex, and propensity for the interaction with the target gene which aggregates scores over all sites on this gene. We also include a list of canonical targets for all considered miRNAs that should be predicted with TargetScan. The current version of the database includes predictions for 2588 human miRNAs and 1915 mouse RNAs. The ncMirTar's webpage includes the ncMirTar webserver for the prediction of targets for novel miRNA sequences in human or mouse genomes (Figure 5-6B). Upon user's query for a given miRNA sequence, the webserver first checks this sequence against all miRNAs from the database and offers a list of the most similar hits. Next, the user has an option to either predict using the original sequence or pick one of the similar miRNAs from the database and retrieve the corresponding results, which is substantially quicker. In either case, the results include the detailed predictions for all non-canonical targets and a list of canonical targets. Moreover, implementation of the ncMirTar predictor can be obtained from http://github.com/BiomineLab/ncMirTar/.



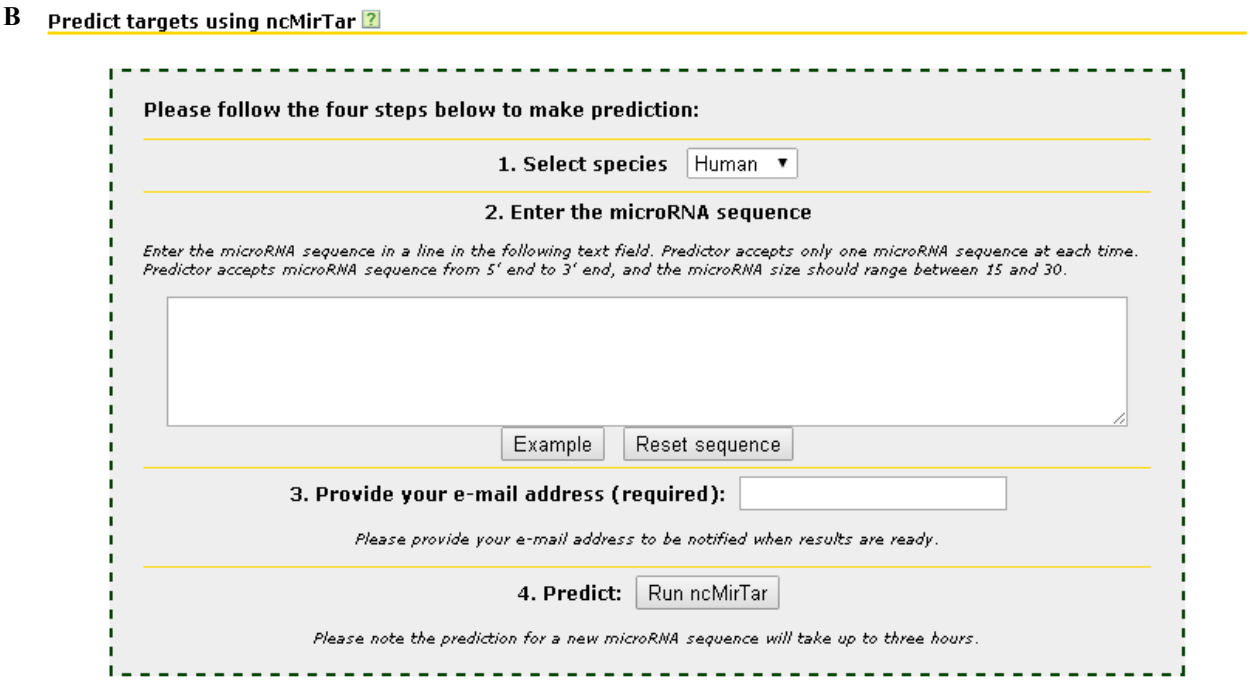


Figure 5-6 Screenshot of ncMirTar webpage
(A) Pre-computed database for searching known miRNA targets. (B) Webserver for predicting targets of new miRNAs.

5.7 Conclusions

We empirically designed, developed, and tested a novel predictive model for high-throughput identification of the non-canonical miRNA targets in 3' UTR of mRNAs. Our model predicts the target genes and miRNA binding sites solely from the sequences of the miRNAs and mRNAs and addresses the lack of methods that accurately predict the non-canonical targets.

We assessed the predictions of miRNA targets using benchmark datasets collected utilizing multiple high- and low-throughput experimental methods for the identification of the miRNA-mRNA interactions. We performed tests on the curated test set of miRNA-target gene pairs and genes that are not repressed by a given miRNA, and using the whole transcriptome/proteome. We also assessed predictions based on the curated test set of the sites on the mRNA which were validated to interact with miRNAs and not to interact with miRNAs.

The tests that utilize a comprehensive set of measures reveal that ncMirTar provides accurate predictions of the non-canonical miRNA targets and significantly outperforms nine representative miRNA target predictors. Our method correctly identifies at least twice as many

non-canonical targets compared to the other methods and provides high true positive rates at low false positive rates. Empirical evaluation shows that ncMirTar accurately predicts both target genes and duplexes (sites on mRNAs). Moreover, the targets for which our predictor outputs high propensity values are characterized by higher degree of repression of their expression levels. We also demonstrate that combining the non-canonical target predictions from ncMirTar and the canonical target predictions from TargetScan leads to a substantial increase in the predictive performance when compared to the other predictors.

Our empirical analysis suggests that the strong predictive performance offered by ncMirTar can be attributed to the design of our predictor. We use a carefully selected training dataset that includes a large population of the non-canonical targets, novel type of features that consider ranking of seed types, and architecture that combines two empirically crafted predictive models for the 6mer and 5mer targets.

To sum up, our predictor offers accurate predictions that complement the current methods that focus on the prediction of the canonical targets. ncMirTar is a high-throughput and cost-effective approach to identify miRNA targets and corresponding binding sites for both specific mRNAs and for whole genomes.

Finally, we made our method conveniently available as a webserver and also provide access to a database of fast-to-query pre-computed results.

Chapter 6

Summary and Future Work

This thesis focuses on characteristics and prediction of miRNA targets. We comprehensively reviewed a large set of miRNA target predictors that were developed over the past decade and shed light on their advantages and drawbacks. We also proposed a new target prediction method to address their major disadvantage related to the predictions for the non-canonical targets.

We started our journey in 2011 through a collaborative project that studied the ER stress pathway using miRNA next generation sequencing data. We implemented a novel pipeline to analyze the miRNA sequencing data that finds known and novel miRNAs, measures their expression levels, quantifies differential expression among samples, and predicts targets of the selected miRNAs. When we integrated the existing miRNA target predictors into our pipeline, we found that they provide very different predictions in the context of the number and overlap between putative targets. We used multiple target prediction methods to potentially improve accuracy of the predictions, but this diversity in predictions shook our confidence in these computational results. This experience also encouraged us to review these predictors to help other users to select appropriate tools for their needs and to offer insights for the developers to design better methods. We summarized the current prediction methods from a comprehensive set of perspectives including their scope, usage, methodology, impact and evaluation procedures. We empirically compared them on benchmark datasets that we carefully curated. We also evaluated relation between their predictive performance and the information that is available before the prediction is performed, which is mostly based on certain characteristics of input miRNAs. Subsequently, we extended our analysis and we found that the current predictors substantially rely on WC base pairing in the seed region. This results in their inability to accurately predict a large number (we estimate it at 50%) of targets that have non-canonical interactions with miRNAs. The prevalence of non-canonical miRNA targets and lack of methods that can accurately predict them motivated us to design a new method to tackle this problem. Drawing from previous works and novel ideas we utilized a large set of old and new features (inputs generated from miRNA and mRNA sequences) and empirically selected a subset of high-

quality features that we used to design our novel predictive model. Our method complements the current methods for prediction of canonical miRNA targets and can be conveniently accessed by the end users via a fast searchable database for the already known miRNAs and a webserver-based predictor for targets of new miRNAs.

6.1 Major contributions

The major contributions of this thesis are in the area of miRNA bioinformatics and they include:

- **Analysis and discovery of miRNAs**.
    - We proposed and implemented a computational pipeline for the analysis of miRNA sequencing data.
    - We applied this novel pipeline to contribute to the discovery of a functional role of a specific miRNA in the ER stress pathway.
- **Overview of current computational miRNA target prediction methods.**
    - We reviewed the largest to date number of 38 current miRNA target predictors from multiple perspectives including their scope, usability, popularity/impact and methodology.
    - We pulled together and developed the most comprehensive to date set of measures to empirically evaluate these predictors.
    - We created and published four new benchmark datasets at four levels of annotations (duplex, gene, transcriptome and proteome) which include true non-functional data (sites on mRNAs or mRNAs that do not interact with the given miRNA).
    - We empirically evaluated and compared a relatively large set of seven popular prediction methods including the most recent methods.
    - For the first time, we analyzed relation between the predictive performance and the information that is available before the prediction is performed.
    - We provided arguably useful insights for the developers to design better target predictors and for the end users to select appropriate tools.
- **Development of an accurate predictor of non-canonical miRNA targets.**
    - We quantified the number of non-canonical miRNA targets for the first time based on multiple data sources.

- We found a common drawback of the current predictors, which is not able to accurately predict the non-canonical miRNA targets.
- We design the first accurate and novel predictor, which takes only the sequence of miRNA and mRNA as its inputs, for the hard-to-predict non-canonical miRNA targets.
- We created new carefully curated and collected from multiple reliable data sources training and test datasets that focus on the non-canonical targets.
- We developed a novel design that utilizes two predictive models to perform prediction of the non-canonical miRNA targets. We also empirically demonstrated that use of these two models results in improved predictive performance when compared with the typical use of a single model.
- We invented a novel feature based on seed type ranking and empirically demonstrated that its use boosts the predictive quality.
- We compared predictive performance of our new method with a relatively large set of nine popular target predictors.

6.2 **Major findings**

By analyzing the historical releases of the miRBase database, we found that the number of miRNAs has grown in quadratic fashion over the last decade. This number has spiked particularly after 2005, which coincides with the deployment of the next generation sequencing. We implemented a new pipeline that generates miRNAs from the sequencing data and we found that it was easy to build thanks to availability of various computational tools for processing the RNA data. Therefore, we conclude that the fast growth is a consequence of the development of the high-throughput sequencing and relative easiness to build computational tools to analyze the miRNA sequencing data. This confirms our first thesis statement.

Although the number of miRNAs grows rapidly, the increase in the number of validated miRNA targets falls far behind. So far only about 4% of known miRNAs have experimentally validated targets. Researchers rely on finding miRNA targets to understand how miRNAs function, and computational tools are very useful in this context. Many miRNA target prediction methods have been developed shortly after the first miRNA was discovered. Based on a comprehensive review of 38 current target predictors, we found that these methods have different scope (take different

inputs, generate different outputs, and consider different species) and offer different functionality (are available in different forms and generate different numbers of often different targets). This finding addresses the second thesis statement. We also performed comprehensive assessment of predictive performance of the existing methods. We found that there is no universally superior predictor, that predictive quality varies widely and that although it is relatively good, further improvements can and should be made. These findings provide support for the third thesis statement.

We found that predictive performance of the current methods is poor for the prediction of non-canonical miRNA targets. This is because these methods heavily rely on the high count of WC base pairs in the seed region. Interestingly, our analysis of five datasets coming from different experimental sources revealed that about half of the targets are non-canonical; this validates the fourth thesis statement. It also implies that the current predictors cannot accurately find other half of the targets. We empirically show that these methods can find on average only 7% of non-canonical miRNA targets. Finally, we designed and comparatively tested first-of-kind sequence-based predictor of the non-canonical miRNA targets. Our empirical tests on several datasets demonstrate that the new method outperforms the current approaches and provides accurate prediction of the non-canonical miRNA targets. These results confirm the final, fifth thesis statement.

6.3 **Future work**

Although undoubtedly computational miRNA target predictors are useful and their predictive performance for the canonical miRNA target is relatively good, further improvements can be made in several areas:

- Current methods utilize many different predictive models. In contrast to other areas of bioinformatics, the empirical (knowledge-based) models (excluding our ncMirTar) do not outperform the heuristic models. This could be due to the low quantity of training data, use of artificial training data (randomly generated non-functional targets), and unbalanced nature of the data (low number of non-functional targets). Thus, one of the future aims should be to improve the quality and quantity of the training data.

- Further improvements in predictive quality could be attained by finding and utilizing not yet known characteristics of miRNA-target interactions. For instance, recently Cis-element was employed to connect primary miRNAs to their potential targets [171], and Gene Ontology annotations and protein-protein interaction networks were used to filter target predictions [191]. Also, the CLIP data have been used to annotate functional targeting sites; however, not much effort so far was made to utilize these data as a filter to improve specificity of the current prediction methods [192].
- We emphasize the need to introduce and maintain higher standards in evaluation of predictive performance, as this would provide a clear picture of current state of this field. Similar to our empirical studies, this should include a comprehensive set of measurements, statistical tests, and use of independent (from the training data) benchmark datasets.
- The outputs generated by the future predictors should be expanded to provide more value for the end users. Some of the possible suggestions include providing location of predicted target sites, allowing predicting targets of novel miRNAs, and predicting the strength of the binding with the help of the gene expression data [193].
- Information of tissue-specific interaction based on the presence of miRNAs should be included in the web servers or databases associated with prediction methods. This function will help users to screen miRNAs based on the tissue that they are interested in.
- Lastly, although the high abundance of the non-canonical miRNA targets has been observed in recent years and confirmed in our study, prediction of these targets did not yet receive enough attention. Although our method provides a much needed solution, further work that would increase the accuracy and coverage (including targets with < 5 WC pairs) is needed.

Bibliography

1. 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**(5)**:**843-854.

2. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans.** *Nature* 2000, **403**(6772)**:**901-906.

3. Huang PJ, Liu YC, Lee CC, Lin WC, Gan RR, Lyu PC, Tang P: **DSAP: deep-sequencing small RNA analysis pipeline.** *Nucleic Acids Res* 2010, **38**(Web Server issue)**:**W385-W391.

4. Ronen R, Gan I, Modai S, Sukacheov A, Dror G, Halperin E, Shomron N: **miRNAkey: a software for microRNA deep sequencing analysis.** *Bioinformatics* 2010, **26**(20)**:**2615-2616.

5. Hackenberg M, Rodriguez-Ezpeleta N, Aransay AM: **miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments.** *Nucleic Acids Res* 2011, **39**(Web Server issue)**:**W132-W138.

6. McBride JL, Pitzer MR, Boudreau RL, Dufour B, Hobbs T, Ojeda SR, Davidson BL: **Preclinical safety of RNAi-mediated HTT suppression in the rhesus macaque as a potential therapy for Huntington's disease.** *Mol Ther* 2011, **19**(12)**:**2152-2162.

7. Seyhan AA: **RNAi: a potential new class of therapeutic for human genetic disease.** *Hum Genet* 2011, **130**(5)**:**583-605.

8. Pereira TC, Lopes-Cendes I: **Emerging RNA-based drugs: siRNAs, microRNAs and derivates.** *Cent Nerv Syst Agents Med Chem* 2012, **12**(3)**:**217-232.

9. Mockenhaupt S, Schurmann N, Grimm D: **When cellular networks run out of control: global dysregulation of the RNAi machinery in human pathology and therapy.** *Prog Mol Biol Transl Sci* 2011, **102:**165-242.

10. Wu W: **MicroRNA: potential targets for the development of novel drugs?** *Drugs R D* 2010, **10**(1)**:**1-8.

11. Thorsen SB, Obad S, Jensen NF, Stenvang J, Kauppinen S: **The therapeutic potential of microRNAs in cancer.** *Cancer J* 2012, **18**(3)**:**275-284.

12. van Rooij E, Purcell AL, Levin AA: **Developing MicroRNA Therapeutics.** *Circulation Research* 2012, **110**(3)**:**496-507.

13. Li Z, Rana TM: **Therapeutic targeting of microRNAs: current status and future challenges.** *Nat Rev Drug Discov* 2014, **13**(8)**:**622-638.

14. Schmidt MF: **Drug target miRNAs: chances and challenges.** *Trends Biotechnol* 2014, **32**(11)**:**578-585.

15. Laitala-Leinonen T: **Update on the development of microRNA and siRNA molecules as regulators of cell physiology.** *Recent Pat DNA Gene Seq* 2010, **4**(2)**:**113-121.

16. Aravalli RN: **Development of MicroRNA Therapeutics for Hepatocellular Carcinoma.** *Diagnostics* 2013, **3**(1)**:**170-191.

17. van Rooij E, Purcell AL, Levin AA: **Developing MicroRNA Therapeutics.** *Circ Res* 2012, **110**(3)**:**496-507.

18. Broderick JA, Zamore PD: **MicroRNA therapeutics.** *Gene Ther* 2011, **18**(12)**:**1104-1110.

19. Tagawa H, Ikeda S, Sawada K: **The role of microRNA in the pathogenesis of malignant lymphoma.** *Cancer Sci* 2013, **104**(17)**:**801-809.

20. Yi B, Piazza GA, Su X, Xi Y: **MicroRNA and Cancer Chemoprevention.** *Cancer Prev Res* 2013, **6**(5)**:**401-409.

21. Sassen S, Miska EA, Caldas C: **MicroRNA: implications for cancer.** *Virchows Arch* 2008, **452**(1)**:**1-10.

22. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human MicroRNA Targets.** *PLoS Biol* 2004, **2**(11)**:**e363.

23. 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**(1)**:**15-20.

24. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434**(7031)**:**338-345.

25. Brennecke J, Stark A, Russell RB, Cohen SM: **Principles of MicroRNA-Target Recognition.** *PLoS Biol* 2005, **3**(3)**:**e85.

26. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N: **Combinatorial microRNA target predictions.** *Nat Genet* 2005, **37**(5)**:**495-500.

27. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I: **A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes.** *Cell* 2006, **126**(6)**:**1203-1217.

28. Griffiths-Jones S: **The microRNA Registry.** *Nucleic Acids Res* 2004, **32**(suppl 1)**:**D109-D111.

29. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature.** *Nucleic Acids Res* 2006, **34**(suppl 1)**:**D140-D144.

30. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Res* 2008, **36**(suppl 1)**:**D154-D158.

31. Kozomara A, Griffiths-Jones S: **miRBase: integrating microRNA annotation and deep-sequencing data.** *Nucleic Acids Res* 2011, **39**(suppl 1)**:**D152-D157.

32. Stark A, Brennecke J, Russell RB, Cohen SM: **Identification of Drosophila MicroRNA targets.** *PLoS Biol* 2003, **1**(3)**:**E60.

33. Lewis BP, Shih I, Jones-Rhoades M, Bartel DP, Burge CB: **Prediction of Mammalian MicroRNA Targets.** *Cell* 2003, **115**(7)**:**787-798.

34. Xu N, Zhang L, Meisgen F, Harada M, Heilborn J, Homey B, Grandér D, Ståhle M, Sonkoly E, Pivarcsi A: **MicroRNA-125b Down-regulates Matrix Metallopeptidase 13 and Inhibits Cutaneous Squamous Cell Carcinoma Cell Proliferation, Migration, and Invasion.** *Journal of Biological Chemistry* 2012, **287**(35)**:**29899-29908.

35. Chen H, Lin Y, Chung H, Lang Y, Lin C, Huang J, Wang W, Lin F, Chen Z, Huang H, Shyy JY-, Liang J, Chen R: **miR-103/107 Promote Metastasis of Colorectal Cancer by Targeting the Metastasis Suppressors DAPK and KLF4.** *Cancer Research* 2012, **72**(14)**:**3631-3641.

36. Muratsu-Ikeda S, Nangaku M, Ikeda Y, Tanaka T, Wada T, Inagi R: **Downregulation of miR-205 Modulates Cell Susceptibility to Oxidative and Endoplasmic Reticulum Stresses in Renal Tubular Cells.** *PLoS ONE* 2012, **7**(7)**:**e41462.

37. Chi SW, Hannon GJ, Darnell RB: **An alternative mode of microRNA target recognition.** *Nat Struct Mol Biol* 2012, **19**(3)**:**321-327.

38. Helwak A, Kudla G, Dudnakova T, Tollervey D: **Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding.** *Cell* 2013, **153**(3)**:**654-665.

39. Hausser J, Zavolan M: **Identification and consequences of miRNA-target interactions--beyond repression of gene expression.** *Nat Rev Genet* 2014, **15**(9)**:**599-612.

40. MacRae IJ, Ma E, Zhou M, Robinson CV, Doudna JA: **In vitro reconstitution of the human RISC-loading complex.** *Proc Natl Acad Sci U S A* 2008, **105**(2)**:**512-517.

41. Sen GL, Wehrman TS, Blau HM: **mRNA translation is not a prerequisite for small interfering RNA-mediated mRNA cleavage.** *Differentiation* 2005, **73**(6)**:**287-293.

42. Khvorova A, Reynolds A, Jayasena SD: **Functional siRNAs and miRNAs Exhibit Strand Bias [abstract].** *Cell* 2003; **115:**209-216.

43. Dangwal S, Thum T: **microRNA therapeutics in cardiovascular disease models.** *Annu Rev Pharmacol Toxicol* 2014, **54:**185-203.

44. Abe M, Bonini NM: **MicroRNAs and neurodegeneration: role and impact [abstract].** *Trends Cell Biol* 2013; **23:**30-36.

45. Czech MP, Aouadi M, Tesz GJ: **RNAi-based therapeutic strategies for metabolic disease.** *Nat Rev Endocrinol* 2011, **7**(8)**:**473-484.

46. Wang Z, Rao DD, Senzer N, Nemunaitis J: **RNA interference and cancer therapy.** *Pharm Res* 2011, **28**(12)**:**2983-2995.

47. Janssen HL, Reesink HW, Lawitz EJ, Zeuzem S, Rodriguez-Torres M, Patel K, van der Meer AJ, Patick AK, Chen A, Zhou Y, Persson R, King BD, Kauppinen S, Levin AA, Hodges MR: **Treatment of HCV infection by targeting microRNA.** *N Engl J Med* 2013, **368**(18)**:**1685-1694.

48. Mazeh H, Mizrahi I, Ilyayev N, Halle D, Brucher B, Bilchik A, Protic M, Daumer M, Stojadinovic A, Itzhak A, Nissan A: **The Diagnostic and Prognostic Role of microRNA in Colorectal Cancer - a Comprehensive review.** *J Cancer* 2013, **4**(3)**:**281-295.

49. Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP: **Prediction of Plant MicroRNA Targets.** *Cell* 2002, **110**(4)**:**513-520.

50. Lai EC: **Predicting and validating microRNA targets.** *Genome Biol* 2004, **5**(9)**:**115.

51. Das N: **MicroRNA Targets - How to predict?** *Bioinformation* 2012, **8**(17)**:**841-845.

52. Bhattacharyya SN, Habermacher R, Martine U, Closs EI, Filipowicz W: **Relief of microRNA-Mediated Translational Repression in Human Cells Subjected to Stress.** *Cell* 2006, **125**(6)**:**1111-1124.

53. Didiano D, Hobert O: **Molecular architecture of a miRNA-regulated 3' UTR.** *RNA* 2008, **14**(7)**:**1297-1317.

54. Hoffman Y, Dahary D, Bublik DR, Oren M, Pilpel Y: **The majority of endogenous microRNA targets within Alu elements avoid the microRNA machinery.** *Bioinformatics* 2013, **29**(7)**:**894-902.

55. Thadani R, Tammi M: **MicroTar: predicting microRNA targets from RNA duplexes.** *BMC Bioinformatics* 2006, **7**(Suppl 5)**:**S20.

56. Lekprasert P, Mayhew M, Ohler U: **Assessing the utility of thermodynamic features for microRNA target prediction under relaxed seed and no conservation requirements.** *PLoS One* 2011, **6**(6)**:**e20622.

57. Guo H, Ingolia NT, Weissman JS, Bartel DP: **Mammalian microRNAs predominantly act to decrease target mRNA levels.** *Nature* 2010, **466**(7308)**:**835-840.

58. Saito T, Saetrom P: **Target gene expression levels and competition between transfected and endogenous microRNAs are strong confounding factors in microRNA high-throughput experiments.** *Silence* 2012, **3**(3)**:**1-15.

59. Kozomara A, Griffiths-Jones S: **miRBase: annotating high confidence microRNAs using deep sequencing data.** *Nucleic Acids Res* 2014, **42**(D1)**:**D68-D73.

60. Hsu S, Tseng Y, Shrestha S, Lin Y, Khaleel A, Chou C, Chu C, Huang H, Lin C, Ho S, Jian T, Lin F, Chang T, Weng S, Liao K, Liao I, Liu C, Huang H: **miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions.** *Nucleic Acids Res* 2014, **42**(D1)**:**D78-D85.

61. 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**(suppl 1)**:**D105-D110.

62. Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, Hatzigeorgiou AG: **TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support.** *Nucleic Acids Res* 2011, **40**(Database-Issue)**:**222-229.

63. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y: **miR2Disease: a manually curated database for microRNA deregulation in human disease.** *Nucleic Acids Res* 2009, **37**(Database issue)**:**D98-D104.

64. Xie B, Ding Q, Han H, Wu D: **miRCancer: a microRNA–cancer association database constructed by text mining on literature.** *Bioinformatics* 2013, **29**(5)**:**638-644.

65. Hosmer DW, Lemeshow S, Sturdivant RX: *Applied Logistic Regression:* 3rd ed. Wiley; 2013.

66. Cortes C, Vapnik V: **Support-vector networks.** *Mach Learning* 1995, **20**(3)**:**273-297.

67. Aizerman MA, Braverman EA, Rozonoer L: **Theoretical foundations of the potential function method in pattern recognition learning.** *Autom Remote Control* 1964, **25:**821-837.

68. Anonymous *Proceedings of the Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory: Jul, 1992;* ACM Press; 1992.

69. Weinstein A: **Review: R. Courant and D. Hilbert, Methods of mathematical physics.** In *Volume 60.* Edited by Anonymous Bulletin of American Mathematical Society; 1954:578-579.

70. Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A: **A combined computational-experimental approach predicts human microRNA targets.** *Genes Dev* 2004, **18**(10)**:**1165-1178.

71. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R: **Fast and effective prediction of microRNA/target duplexes.** *RNA* 2004, **10**(10)**:**1507-1517.

72. Grun D, Wang Y, Langenberger D, Gunsalus KC, Rajewsky N: **microRNA Target Predictions across Seven *Drosophila* Species and Comparison to Mammalian Targets.** *PLoS Comput Biol* 2005, **1**(1)**:**e13.

73. Fawcett T: **An introduction to ROC analysis.** *Pattern Recogn Lett* 2006, **27**(8)**:**861-874.

74. Marín RM, Vaníček J: **Efficient use of accessibility in microRNA target prediction.** *Nucleic Acids Res* 2011, **39**(1)**:**19-29.

75. Liu H, Yue D, Chen Y, Gao SJ, Huang Y: **Improving performance of mammalian microRNA target prediction.** *BMC Bioinformatics* 2010, **11:**476.

76. Chandra V, Girijadevi R, Nair A, Pillai S, Pillai R: **MTar: a computational microRNA target prediction architecture for human transcriptome.** *BMC Bioinformatics* 2010, **11:**S2.

77. Sturm M, Hackenberg M, Langenberger D, Frishman D: **TargetSpy: a supervised machine learning approach for microRNA target prediction.** *BMC Bioinformatics* 2010, **11:**292.

78. Bandyopadhyay S, Mitra R: **TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples.** *Bioinformatics* 2009, **25**(20)**:**2625-2631.

79. Ruan J, Chen H, Kurgan L, Chen K, Kang C, Pu P: **HuMiTar: a sequence-based method for prediction of human microRNA targets.** *Algorithms Mol Biol* 2008, **3:**16.

80. Anonymous *Proceedings of the Proceedings of the 14th International Joint Conference on Artificial Intelligence:* Morgan Kaufmann; 1995.

81. Box JF: **Guinness, Gosset, Fisher, and Small Samples.** *Statist Sci* 1987, **2**(1)**:**45-52.

82. Fay MP, Proschan MA: **Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules.** *Stat Surv* 2010, **4:**1-39.

83. Anderson TW, Darling DA: **Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes.** *Ann Math Statist* 1952, **23**(2)**:**193-212.

84. Wu R: **Nucleotide sequence analysis of DNA: I. Partial sequence of the cohesive ends of bacteriophage λ and 186 DNA.** *J Mol Biol* 1970, **51**(3)**:**501-521.

85. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437**(7057)**:**376-380.

86. Groenendyk J, Peng Z, Dudek E, Fan X, Mizianty MJ, Dufey E, Urra H, Sepulveda D, Rojas-Rivera D, Lim Y, Kim do H, Baretta K, Srikanth S, Gwack Y, Ahnn J, Kaufman RJ, Lee SK, Hetz C, Kurgan L, Michalak M: **Interplay between the oxidoreductase PDIA6 and**

microRNA-322 controls the response to disrupted endoplasmic reticulum calcium homeostasis. *Sci Signal* 2014, **7**(329)**:**ra54.

87. Groenendyk J, Fan X, Peng Z, Ilnytskyy Y, Kurgan L, Michalak M: **Genome-wide analysis of thapsigargin-induced microRNAs and their targets in NIH3T3 cells.** *Genomics Data* 2014, **2:**325-327.

88. Sun Z, Evans J, Bhagwate A, Middha S, Bockol M, Yan H, Kocher J: **CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data.** *BMC Genomics* 2014, **15**(1)**:**423.

89. Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ: **Kraken: A set of tools for quality control and analysis of high-throughput sequence data.** *Methods* 2013, **63**(1)**:**41-49.

90. Zhang Y, Xu B, Yang Y, Ban R, Zhang H, Jiang X, Cooke HJ, Xue Y, Shi Q: **CPSS: a computational platform for the analysis of small RNA deep sequencing data.** *Bioinformatics* 2012, **28**(14)**:**1925-1927.

91. Wen M, Shen Y, Shi S, Tang T: **miREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments.** *BMC Bioinformatics* 2012, **13:**140.

92. Fasold M, Langenberger D, Binder H, Stadler PF, Hoffmann S: **DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments.** *Nucleic Acids Res* 2011, **39**(Web Server issue)**:**W112-W117.

93. Zhao W, Liu W, Tian D, Tang B, Wang Y, Yu C, Li R, Ling Y, Wu J, Song S, Hu S: **wapRNA: a web-based application for the processing of RNA sequences.** *Bioinformatics* 2011, **27**(21)**:**3076-3077.

94. Yang J, Shao P, Zhou H, Chen Y, Qu L: **deepBase: a database for deeply annotating and mining deep sequencing data.** *Nucleic Acids Res* 2010, **38**(suppl 1)**:**D123-D130.

95. Mathelier A, Carbone A: **MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data.** *Bioinformatics* 2010, **26**(18)**:**2226-2234.

96. Pantano L, Estivill X, Marti E: **SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells.** *Nucleic Acids Res* 2010, **38**(5)**:**e34.

97. Sales G, Coppe A, Bisognin A, Biasiolo M, Bortoluzzi S, Romualdi C: **MAGIA, a web-based tool for miRNA and Genes Integrated Analysis.** *Nucleic Acids Res* 2010, **38**(Web Server issue)**:**W352-W359.

98. Zhu E, Zhao F, Xu G, Hou H, Zhou L, Li X, Sun Z, Wu J: **mirTools: microRNA profiling and discovery based on high-throughput sequencing.** *Nucleic Acids Res* 2010, **38**(suppl 2)**:**W392-W397.

99. Wang WC, Lin FM, Chang WC, Lin KY, Huang HD, Lin NS: **miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression.** *BMC Bioinformatics* 2009, **10:**328.

100. Hackenberg M, Sturm M, Langenberger D, Falcon-Perez JM, Aransay AM: **miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments.** *Nucleic Acids Res* 2009, **37**(Web Server issue)**:**W68-W76.

101. Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N: **Discovering microRNAs from deep sequencing data using miRDeep.** *Nat Biotechnol* 2008, **26**(4)**:**407-415.

102. Kong Y: **Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies.** *Genomics* 2011, **98**(2)**:**152-153.

103. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**(3)**:**R25-2009-10-3-r25. Epub 2009 Mar 4.

104. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110**(1-4)**:**462-467.

105. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucleic Acids Res* 2005, **33**(Database issue)**:**D121-D124.

106. Hofacker IL: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31**(13)**:**3429-3431.

107. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26**(1)**:**139-140.

108. Tan Gana NH, Victoriano AF, Okamoto T: **Evaluation of online miRNA resources for biomedical applications.** *Genes Cells* 2012, **17**(1)**:**11-27.

109. Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Vergoulis T, Koziris N, Sellis T, Tsanakas P, Hatzigeorgiou AG: **DIANA-microT web server: elucidating microRNA functions through target prediction.** *Nucleic Acids Res* 2009, **37**(suppl 2)**:**W273-W276.

110. Elefant N, Berger A, Shein H, Hofree M, Margalit H, Altuvia Y: **RepTar: a database of predicted cellular targets of host and viral miRNAs.** *Nucleic Acids Res* 2011, **39**(Database issue)**:**D188-D194.

111. Fan X, Kurgan L: **Comprehensive overview and assessment of computational prediction of microRNA targets in animals.** *Brief Bioinform* 2014, **:**.

112. Pio G, Malerba D, D'Elia D, Ceci M: **Integrating microRNA target predictions for the discovery of gene regulatory networks: a semi-supervised ensemble learning approach.** *BMC Bioinformatics* 2014, **15**(Suppl 1)**:**S4.

113. Corrada D, Viti F, Merelli I, Battaglia C, Milanesi L: **myMIR: a genome-wide microRNA targets identification and annotation tool.** *Brief Bioinform* 2011, **12**(6)**:**588-600.

114. Rajewsky N, Socci ND: **Computational identification of microRNA targets.** *Dev Biol* 2004, **267**(2)**:**529-535.

115. Sethupathy P, Megraw M, Hatzigeorgiou AG: **A guide through present computational approaches for the identification of mammalian microRNA targets.** *Nat Methods* 2006, **3**(11)**:**881-886.

116. Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG: **Lost in translation: an assessment and perspective for computational microRNA target identification.** *Bioinformatics* 2009, **25**(23)**:**3049-3055.

117. Peterson SM, Thompson JA, Ufkin ML, Sathyanarayana P, Liaw L, Congdon CB: **Common features of microRNA target prediction tools.** *Front Genet* 2014, **5:**23.

118. Reyes-Herrera PH, Ficarra E: **One Decade of Development and Evolution of MicroRNA Target Prediction Algorithms.** *Genomics, Proteomics Bioinformatics* 2012, **10**(5)**:**254-263.

119. Witkos TM, Koscianska E, Krzyzosiak WJ: **Practical Aspects of microRNA Target Prediction.** *Curr Mol Med* 2011, **11**(2)**:**93-109.

120. Saito T, Sætrom P: **MicroRNAs – targeting and target prediction.** *New Biotechnol* 2010, **27**(3)**:**243-249.

121. Min H, Yoon S: **Got target? Computational methods for microRNA target prediction and their extension.** *Exp Mol Med* 2010, **42**(4)**:**233-244.

122. Li L, Xu J, Yang D, Tan X, Wang H: **Computational approaches for microRNA studies: a review.** *Mamm Genome* 2010, **21**(1-2)**:**1-12.

123. Yousef M, Showe L, Showe M: **A study of microRNAs in silico and in vivo: bioinformatics approaches to microRNA discovery and target identification.** *FEBS J* 2009, **276**(8)**:**2150-2156.

124. Mendes ND, Freitas AT, Sagot MF: **Current tools for the identification of miRNA genes and their targets.** *Nucleic Acids Res* 2009, **37**(8)**:**2419-2433.

125. Bartel DP: **MicroRNAs: target recognition and regulatory functions.** *Cell* 2009, **136**(2)**:**215-233.

126. Watanabe Y, Tomita M, Kanai A: **Computational methods for microRNA target prediction.** *Methods Enzymol* 2007, **427:**65-86.

127. Maziere P, Enright AJ: **Prediction of microRNA targets.** *Drug Discov Today* 2007, **12**(11-12)**:**452-458.

128. Chaudhuri K, Chatterjee R: **MicroRNA detection and target prediction: integration of computational and experimental approaches.** *DNA Cell Biol* 2007, **26**(5)**:**321-337.

129. Zhang B, Pan X, Wang Q, Cobb GP, Anderson TA: **Computational identification of microRNAs and their targets.** *Comput Biol Chem* 2006, **30**(6)**:**395-407.

130. Rajewsky N: **microRNA target predictions in animals.** *Nat Genet* 2006, **38**(Suppl)**:**S8-S13.

131. Srivastava P, Moturu T, Pandey P, Baldwin I, Pandey S: **A comparison of performance of plant miRNA target prediction tools and the characterization of features for genome-wide target prediction.** *BMC Genomics* 2014, **15**(1)**:**348.

132. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N,

Davis S, Soboleva A: **NCBI GEO: archive for functional genomics data sets—update.** *Nucleic Acids Res* 2013, **41**(D1)**:**D991-D995.

133. Schwanhausser B, Gossen M, Dittmar G, Selbach M: **Global analysis of cellular protein translation by pulsed SILAC.** *Proteomics* 2009, **9**(1)**:**205-209.

134. Choudhuri S: **Small noncoding RNAs: biogenesis, function, and emerging significance in toxicology.** *J Biochem Mol Toxicol* 2010, **24**(3)**:**195-216.

135. Sun BK, Tsao H: **Small RNAs in development and disease.** *J Am Acad Dermatol* 2008, **59**(5)**:**738-740.

136. Sood P, Krek A, Zavolan M, Macino G, Rajewsky N: **Cell-type-specific signatures of microRNAs on target mRNA expression.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(8)**:**2746-2751.

137. Guo Z, Maki M, Ding R, Yang Y, zhang B, Xiong L: **Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues.** *Scientific Reports* 2014, **4:**5150.

138. Londin E, Loher P, Telonis AG, Quann K, Clark P, Jing Y, Hatzimichael E, Kirino Y, Honda S, Lally M, Ramratnam B, Comstock CES, Knudsen KE, Gomella L, Spaeth GL, Hark L, Katz LJ, Witkiewicz A, Rostami A, Jimenez SA, Hollingsworth MA, Yeh JJ, Shaw CA, McKenzie SE, Bray P, Nelson PT, Zupo S, Van Roosbroeck K, Keating MJ, Calin GA, Yeo C, Jimbo M, Cozzitorto J, Brody JR, Delgrosso K, Mattick JS, Fortina P, Rigoutsos I: **Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs.** *Proceedings of the National Academy of Sciences* 2015, **112**(10)**:**E1106-E1115.

139. BABAK T, ZHANG W, MORRIS Q, BLENCOWE BJ, HUGHES TR: **Probing microRNAs with microarrays: Tissue specificity and functional inference.** *RNA* 2004, **10**(11)**:**1813-1819.

140. Hsieh WJ, Lin F, Huang H, Wang H: **Investigating microRNA-Target Interaction-Supported Tissues in Human Cancer Tissues Based on miRNA and Target Gene Expression Profiling.** *PLoS ONE* 2014, **9**(4)**:**e95697.

141. Linsley PS, Schelter J, Burchard J, Kibukawa M, Martin MM, Bartz SR, Johnson JM, Cummins JM, Raymond CK, Dai H, Chau N, Cleary M, Jackson AL, Carleton M, Lim L: **Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression.** *Mol Cell Biol* 2007, **27**(6)**:**2240-2252.

142. He L, He X, Lim LP, de Stanchina E, Xuan Z, Liang Y, Xue W, Zender L, Magnus J, Ridzon D, Jackson AL, Linsley PS, Chen C, Lowe SW, Cleary MA, Hannon GJ: **A microRNA component of the p53 tumour suppressor network.** *Nature* 2007, **447**(7148)**:**1130-1134.

143. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP: **MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing.** *Mol Cell* 2007, **27**(1)**:**91-105.

144. 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**(8)**:**R90.

145. Vejnar CE, Zdobnov EM: **miRmap: Comprehensive prediction of microRNA target repression strength.** *Nucleic Acids Res* 2012, **40**(22)**:**11673-11683.

146. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N: **Widespread changes in protein synthesis induced by microRNAs.** *Nature* 2008, **455**(7209)**:**58-63.

147. Wang X, El Naqa IM: **Prediction of both conserved and nonconserved microRNA targets in animals.** *Bioinformatics* 2008, **24**(3)**:**325-332.

148. Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM: **Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution.** *Cell* 2005, **123**(6)**:**1133-1146.

149. Friedman RC, Farh KK, Burge CB, Bartel D: **Most mammalian mRNAs are conserved targets of microRNAs.** *Genome Res* 2008, **19**(1)**:**92-105.

150. Maragkakis M, Alexiou P, Papadopoulos GL, Reczko M, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Simossis VA, Sethupathy P, Vergoulis T, Koziris N, Sellis T, Tsanakas P, Hatzigeorgiou AG: **Accurate microRNA target prediction correlates with protein repression levels.** *BMC Bioinformatics* 2009, **10:**295.

151. Maragkakis M, Vergoulis T, Alexiou P, Reczko M, Plomaritou K, Gousis M, Kourtis K, Koziris N, Dalamagas T, Hatzigeorgiou AG: **DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association.** *Nucleic Acids Res* 2011, **39**(Web Server issue)**:**W145-W148.

152. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG: **DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows.** *Nucleic Acids Res* 2013, **41**(Web Server issue)**:**W169-W173.

153. Anders G, Mackowiak SD, Jens M, Maaskola J, Kuntzagk A, Rajewsky N, Landthaler M, Dieterich C: **doRiNA: a database of RNA interactions in post-transcriptional regulation.** *Nucleic Acids Res* 2012, **40**(Database issue)**:**D180-D186.

154. Kurgan L, Disfani FM: **Structural protein descriptors in 1-dimension and their sequence-based predictions.** *Curr Protein Pept Sci* 2011, **12**(6)**:**470-489.

155. de Ridder D, de Ridder J, Reinders MJT: **Pattern recognition in bioinformatics.** *Brief Bioinform* 2013, **14**(5)**:**633-647.

156. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS: **MicroRNA targets in Drosophila.** *Genome Biol* 2003, **5**(1)**:**R1.

157. Robins H, Li Y, Padgett RW: **Incorporating structure to predict microRNA targets.** *Proc Nat Acad Sci USA* 2005, **102**(11)**:**4006-4009.

158. Burgler C, Macdonald PM: **Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method.** *BMC Genomics* 2005, **6:**88.

159. Rusinov V, Baev V, Minkov IN, Tabler M: **MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence.** *Nucleic Acids Res* 2005, **33**(Web Server issue)**:**W696-W700.

160. Saetrom O, Snove O,Jr, Saetrom P: **Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms.** *RNA* 2005, **11**(7)**:**995-1003.

161. Kim SK, Nam JW, Rhee JK, Lee WJ, Zhang BT: **miTarget: microRNA target gene prediction using a support vector machine.** *BMC Bioinformatics* 2006, **7:**411.

162. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M: **Inference of miRNA targets using evolutionary conservation and pathway analysis.** *BMC Bioinformatics* 2007, **8:**69.

163. Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y: **Potent effect of target structure on microRNA function.** *Nat Struct Mol Biol* 2007, **14**(4)**:**287-294.

164. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E: **The role of site accessibility in microRNA target recognition.** *Nat Genet* 2007, **39**(10)**:**1278-1284.

165. Nielsen C, Shomron N, Sandberg R, Hornstein E, Kitzman J, Burge C: **Determinants of targeting by endogenous and exogenous microRNAs and siRNAs.** *RNA* 2007, **13**(11)**:**1894-1910.

166. Mitra R, Bandyopadhyay S: **MultiMiTar: a novel multi objective optimization based miRNA-target prediction method.** *PLoS ONE* 2011, **6**(9)**:**e24583.

167. Reyes Herrera PH, Ficarra E, Acquaviva A, Macii E: **miREE: miRNA Recognition Elements Ensemble.** *BMC Bioinformatics* 2011, **12**(1)**:**454.

168. Jeggari A, Marks DS, Larsson E: **miRcode: a map of putative microRNA target sites in the long non-coding transcriptome.** *Bioinformatics* 2012, **28**(15)**:**2062-2063.

169. Ahmadi H, Ahmadi A, Azimzadeh-Jamalkandi S, Shoorehdeli MA, Salehzadeh-Yazdi A, Bidkhori G, Masoudi-Nejad A: **HomoTarget: A new algorithm for prediction of microRNA targets in Homo sapiens.** *Genomics* 2013, **101**(2)**:**94-100.

170. Liu H, Zhou S, Guan J: **Identifying Mammalian MicroRNA Targets Based on Supervised Distance Metric Learning.** *IEEE Trans Inf Technol Biomed* 2012, **17**(2)**:**427-435.

171. Fujiwara T, Yada T: **miRNA-target prediction based on transcriptional regulation.** *BMC Genomics* 2013, **14**(Suppl 2)**:**S3.

172. Hoshi T, Zagotta WN, Aldrich RW: **Biophysical and molecular mechanisms of Shaker potassium channel inactivation.** *Science* 1990, **250**(4980)**:**533-538.

173. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31**(13)**:**3406-3415.

174. Markham NR, Zuker M: **DINAMelt web server for nucleic acid melting prediction.** *Nucleic Acids Res* 2005, **33**(Web Server issue)**:**W577-W581.

175. Ding Y, Chan CY, Lawrence CE: **Sfold web server for statistical folding and rational design of nucleic acids.** *Nucleic Acids Res* 2004, **32**(suppl 2)**:**W135-W141.

176. Vasudevan S, Steitz JA: **AU-Rich-Element-Mediated Upregulation of Translation by FXR1 and Argonaute 2.** *Cell* 2007, **128**(6)**:**1105-1118.

177. Hausser J, Syed AP, Bilen B, Zavolan M: **Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation.** *Genome Res* 2013, **23**(4)**:**604-615.

178. Marín RM, Šulc M, Vaníček J: **Searching the coding region for microRNA targets.** *RNA* 2013, **19**(4)**:**467-474.

179. Sethupathy P, Corda B, Hatzigeorgiou AG: **TarBase: A comprehensive database of experimentally supported animal microRNA targets.** *RNA* 2006, **12**(2)**:**192-197.

180. Betel D, Wilson M, Gabow A, Marks DS, Sander C: **The microRNA.org resource: targets and expression.** *Nucleic Acids Res* 2008, **36**(Database issue)**:**D149-D153.

181. Fang Z, Rajewsky N: **The impact of miRNA target sites in coding sequences and in 3'UTRs.** *PLoS One* 2011, **6**(3)**:**e18067.

182. Doench JG, Sharp PA: **Specificity of microRNA target selection in translational repression.** *Gene Dev* 2004, **18**(5)**:**504-511.

183. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP: **Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs.** *Nat Struct Mol Biol* 2011, **18**(10)**:**1139-1146.

184. Kim KK, Ham J, Chi SW: **miRTCat: a comprehensive map of human and mouse microRNA target sites including non-canonical nucleation bulges.** *Bioinformatics* 2013, **29**(15)**:**1898-1899.

185. Khorshid M, Hausser J, Zavolan M, van Nimwegen E: **A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets.** *Nat Methods* 2013, **10**(3)**:**253-255.

186. Gumienny R, Zavolan M: **Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G.** *Nucleic Acids Res* 2015, **43**(3)**:**1380-1391.

187. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives.** *Nucleic Acids Res* 2009, **37**(Database issue)**:**D32-D36.

188. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl Toolkit: Perl Modules for the Life Sciences.** *Genome Res* 2002, **12**(10)**:**1611-1618.

189. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**(1)**:**195-197.

190. Chang C, Lin C: **LIBSVM: A Library for Support Vector Machines.** *ACM Trans Intell Syst Technol* 2011, **2**(3)**:**27.

191. Wang P, Ning S, Wang Q, Li R, Ye J, Zhao Z, Li Y, Huang T, Li X: **mirTarPri: improved prioritization of microRNA targets through incorporation of functional genomics data.** *PLoS One* 2013, **8**(1)**:**e53685.

192. Li J, Liu S, Zhou H, Qu L, Yang J: **starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data.** *Nucleic Acids Res* 2014, **42**(D1)**:**D92-D97.

193. Radfar H, Wong W, Morris Q: **BayMiR: inferring evidence for endogenous miRNA-induced gene repression from mRNA expression profiles.** *BMC Genomics* 2013, **14**(1)**:**592.