Send Orders of Reprints at reprints@benthamscience.org

The Open Biology Journal, 2012, *5*, 23-30

Open Access

Computational Tools for Genome-Wide miRNA Prediction and Study

Tareq B. Malas[1] and Timothy Ravasi[2,*]

[1]*Division of Biological and Environmental Sciences & Engineering, Division of Applied Mathematics and Computer Sciences, Computational Biosciences Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia*

[2]*King Abdullah University of Science and Technology Thuwal 23955 Kingdom of Saudi Arabia. Division of Medical Genetics, Department of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 USA*

Abstract: MicroRNAs (miRNAs) are single-stranded non-coding RNA susually of 22 nucleotidesin length that play an important post-transcriptional regulation role in many organisms. MicroRNAs bind a seed sequence to the 3′-untranslated region (UTR) region of the target messenger RNA (mRNA), inducing degradation or inhibition of translation and resulting in a reduction in the protein level. This regulatory mechanism is central to many biological processes and perturbation could lead to diseases such as cancer. Given the biological importance, of miRNAs, there is a great need to identify and study their targets and functions. However, miRNAs are very difficult to clone in the lab and this has hindered the identification of novel miRNAs. Next-generation sequencing coupled with new computational tools has recently evolved to help researchers efficiently identify large numbers of novel miRNAs. In this review, we describe recent miRNA prediction tools and discuss their priorities, advantages and disadvantages.

Keywords: miRNAs, Computational prediction tools, Gene regulation, Biological databases, Genomics.

INTRODUCTION

MicroRNAs (miRNAs) are single-stranded non-coding RNAsthat are approximately 22 nucleotides in length. miRNAs are important post-transcriptional regulators that have been shown to play fundamentally important regulatory roles in animal and plant development and to be involved in diverse cellular and physiological events such as apoptosis, proliferation, tumorigenes is and genetic disorders. miRNAs in mammals bind a complementary base pair to the 3′-untranslated region (3′ UTR) of the target mRNA, causing the inhibition of translation and/or degradation of that mRNA [1].

miRNA coding sequences can be found in introns or exons of protein-coding sequences or in intergenic regions [2]. miRNA genes are transcribed by RNA polymerase II to produce primary-miRNAs (pri-mRNA) [3]. pri-miRNAs are several kilo bases long and contain a local hairpin structure, which iscleaved by the nuclear RNase III Drosha that worksin complexes with dsRNA-binding proteins to form the Drosha-DGCR8 complex in the nucleus to release the stem-looped hairpin precursor miRNA (pre-miRNA). The resulting pre-miRNA is exported to the cytoplasm through nuclear pore complexes mediated by the nuclear transport receptor, Exporting-5, and is then further processed by another RNase III enzyme called Dicer to finally generate the 22-nucleotide mature miRNA-miRNA duplex [4]. Not all miRNAs need to be cleaved by Drosha, where a distinct class of miRNAs termed *mirtrons* are directly transcribed into pre-miRNAs from introns of similar sizes and exported to the cytoplasm for Dicer processing [5]. At this stage, only one strand of the mature miRNA is retained and incorporated into the miRNA-induced silencing complex (miRISC) while the other is degraded. In addition to the single-stranded mature miRNA, the miRISC complex includes members of the Ago protein family. The Ago proteins contain two conserved RNA binding domains that bind the mature miRNA and orient it for the interaction with target mRNA and play a critical role in miRNA-induced silencing [6] (Fig. **1**).

A recent study estimates the number of miRNAs in the human genome to be about 55,000 [7], which is much greater than the experimentally verified miRNAs published in the literature, and it is estimated that over 30% of protein coding genes in *Homo sapiens* are regulated by miRNA [8].Currently 1,527 *Homo sapiens* miRNAs are listed inmiR Base version 18.0 [9]. This comparatively small number of experimentally verified miRNAs is not surprising because miRNAs are difficult to clone and detect in the lab given their short lengths, low expression levels and selective expression patterns (e.g., they exhibit highly constrained tissue- and time-specific patterns) [10]. Discovery of new miRNAs and a better understanding of their biological functions in general and in oncology and disease-development in particular will further enable scientists to use miRNAs as drug targets. Santaris Pharma is such a company that is focused on the study and development of RNA-targeted therapies, has announced the development of the first micro RNA-targeted drug to enter clinical trials *miravirsen*, which is used to treat patients infected with Hepatitis C virus [11].

In silico miRNA prediction tools are being developed to overcome the limitations of identifying novel miRNAs in the

*Address correspondence to this author at the King Abdullah University of Science and Technology Thuwal 23955 Kingdom of Saudi Arabia. Division of Medical Genetics, Department of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 USA; Tel: +966-5-44700067, Fax: +966-2-8020127; E-mail: timothy.ravasi@kaust.edu.sa

1874-1967/12

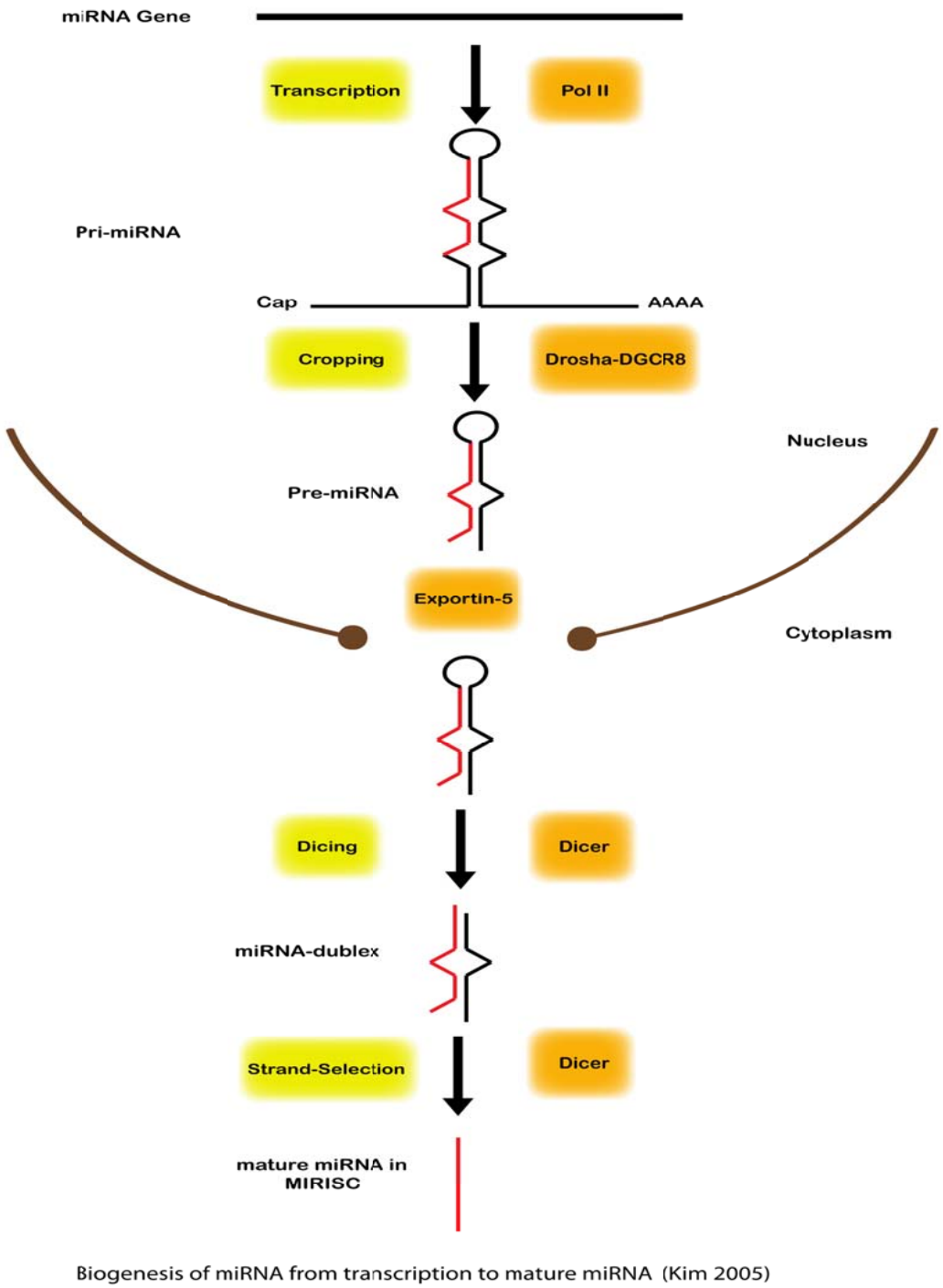

Fig. (1). Biogenesis of miRNA from transcription to maturity (Kim 2005).

lab. Most of these tools seek to identify sequences that qualify as pre-miRNAs by focusing on their high conservation and unique biological and structural features. *In silico* miRNA prediction has gained traction as the first step towards identifying novel miRNAs and has helped scientists identify defined sets of candidates to test experimentally. Current miRNA prediction tools are able to predict novel miRNAs with reasonable success, yet there is a much room for improvement and it is expected that *in silico* miRNA predictiontools will continually improve. With the ever increasing number of *in silico* miRNA prediction and study tools, there is a strong need to classify and organize these tools into functional categories and compare their performances in an attempt to help researchers in selecting the tool most suitable for their study. Here, we review existing miRNA prediction tools and classify them into comparative and non-comparative tools based on their prediction methods (other sorts of classifications can be seen in [12, 13]) and cover the emerging field of identifying miRNAs *via* deep sequencing.

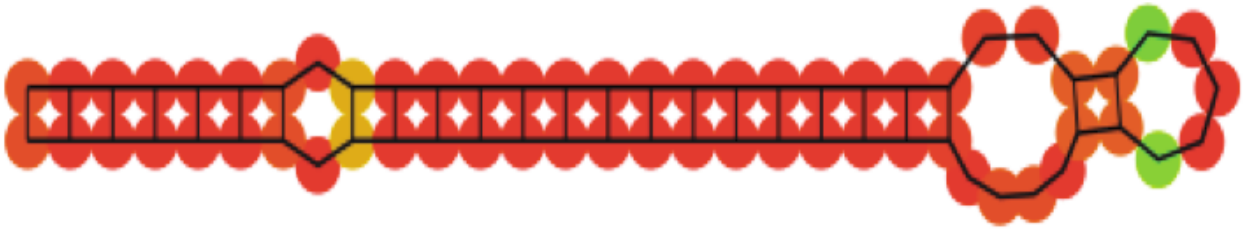

Fig. (2). Human pre-miRNA has an mir-520b hairpin secondary structure predicted by the RNAfold program (Hofacker, Fontana . 1994).

Table 1. Examples of Common miRNA Features Used for *in silico* miRNA Prediction

Feature Type	Characteristic	Example
Sequence	G+C content; dinucleotide frequencies	DIANA-microH; miPred
Conservation	Sequence alignment across species; arm conservation	RNAmicro; DIANA-microH
Thermodynamical	Free energy of folding; minimum folding energy (MFE)	RNAmicro; miPred
Structure	Number of nucleotides in symmetrical and asymmetrical loops; loop and stem sizes	miR-abel; miRNA SVM

FEATURES OF A MIRNA

miRNAs possess unique characteristics in terms of sequence and structure that are functionally important in performing their tasks as post-transcriptional regulators. Accurate *in silico* miRNA prediction requires computational tools to identify these specific characteristics to distinguish real miRNAs from pseudo ones. The most important of these characteristics is that of the pre-miRNA, which should possess a statistically significant and evolutionarily conserved (a) symmetric RNA hairpin (Fig. **2**). In addition, the adjusted minimum free energy level of the pre-miRNA stem-loop structure should be low in order for it to be stable [14]. This hairpin structure is important during mature miRNA biogenesis where it acts as a structural motif for Exportin-5 in nuclear-cytoplasm transportation and later as a substrate for Dicer (Fig. **1**) [15]. The hairpin feature should be distinct from those of random inverted repeats that can fold into a dysfunctional hairpin (pseudo miRNAs) and from other non-coding (nc) RNAs that are capable of forming hairpin-like structures [16]. Another characteristic of most miRNAs is their profound conservation across closely related species [1]. In general, miRNA prediction tools rely on these characteristics and other sequence, structure, conservation and thermodynamical features, which they are able to extract from a set of already experimentally verified miRNAs to conduct their predictions and enhance their accuracy (Table **1**). The discovery of new experimentally verified miRNAs will help miRNA prediction tools extract new miRNA-specific features and characteristics that can be used to improve the prediction accuracy of these tools.

COMPUTATIONAL TOOLS FOR GENOMIC PREDICTION AND DISCOVERY OF MIRNAS

Comparative Tools

Comparative tools rely on miRNA conservation acrossspecies to identify novel ones. The rationale behind comparative methods is to identify genome sequences that can fold into hairpin-like structures and become conserved among species as pre-miRNA candidates. Earlier attempts by Batuwita and Palade relied on identifying close homologs of published pre-miRNA, e.g., let-7 [17]. This can be as straightforward as using BLAST search to identify the homology between sequences and then testing the candidates for their ability to fold into hairpin-like structures [16]. Tools in this category utilize machine-learning algorithms such as Support Vector Machines (SVM) to evaluate pre-miRNA characteristics in addition to relying on conservation and homology. The main advantage of such tools is their ability to discover well-conserved, genome-wide pre-miRNAs, although they clearly lack the ability to discover novel miRNAs that lack clear homologues. It is worth noting that the human genome contains a large number of sequences that can fold into hairpin-like structures, most of which are pseudo hairpins that have a variety of functions and are conserved among different species [18]. In addition to these conserved pseudo hairpins, other types of ncRNAs can have motifs capable of folding into hairpin-like structures [19]. Thus, for comparative prediction tools to work effectively, they should be able to distinguish between a real pre-miRNA, a conserved pseudo hairpin and other types of ncRNAs that can fold into hairpin-like structures [20]. Among the available comparative tools, only RNA micro [21] and DIANA-microH [22] partially considered this issue. RNAmicro considers the classification of real pre-miRNAs from other types of ncRNAs that fold into hairpin-like structures, and DIANA-microH differentiates pre-miRNAs from pseudo hairpins.

Non-Comparative Tools

Non-comparative tools use computational recognition techniques (e.g., machine learning) to distinguish between real pre-miRNAs and pseudo hairpins. They do notrely on phylogenetic conservation signals and are thus able to predict novel non-conserved/species-specific miRNAs [20]. Non-comparative tools are becoming the trend amongmi RNA prediction methods and the majority of the recently developed tools fall into this category. Most of the non-comparative tools start by defining unique features (e.g., structure, sequence) of miRNAs to use as the basis for distinguishing between real pre-miRNAs and pseudo ones. Most of these tools rely on a set of positive and negative pre-miRNAs from which to extract these features; these datasets

Table 2. Performance Comparison Between Different Non-Comparative Tools

Tool Name	Classifier	Sensitivity	Specificity	Accuracy
SSCprofiler [23]	Profile HMM	88.95%	84.16%	72.15%
miPred [16]	SVM	84.55%	97.97%	93.50%
microPred [20]	SVM	90.20%	97.28%	NA
miRD [24]	SVM;Boosting	NA	NA	94.0%

are termed training datasets. One of the most challenging tasks in constructing a non-comparative miRNA prediction model is in selecting high-quality training datasets. The positive dataset must be inclusive of all types of pre-miRNAs in order to reduce false negatives as much as possible and the negative dataset must include pseudo miRNAs that are similar in characteristics to real miRNAs but different enough that the model will reduce false positives as much as possible. Over the past few years, different machine learning algorithms were used to predict novel miRNAs with high accuracy. SSC profiler, a Profile Hidden Markov Model(HMM)-based prediction model designed by Oulas *et al*., [23] used 249 human sequences of experimentally validated pre-miRNAs as a training set to extract sequence, structure and conservation feature characteristics of pre-miRNAs. These extracted features were then applied to genomic regions to identify novel miRNA precursors. Their negative dataset included 35,000 sequences generated from 3′-UTR regions. The authors reasoned that since no experimentally verified miRNA hadnever been found in 3′-UTRs, they could generate a negative set of pseudo hairpins from these regions with high certainty of not including any real miRNA in the set while still sharing biological characteristics. To test their model, they screened 350MB of Cancer Associated Genomic Regions (CAGRs) for novel miRNAs. This screening resulted in identifying 20 pre-miRNA candidates that wereexpressed in the HeLa cell-line.

Support Vector Machine (SVM) is also being employed as a model for computational prediction tools. MiPred is anSVM-based prediction tool published by Ng and Mishra [16]. MiPred utilizes 23 global and intrinsic features of pre-miRNA folding measures to distinguish true miRNA precursors from pseudo hairpins. The authors used 200 experimentally verified human miRNA precursors and 400 pseudo hairpins to train their model. However, they failed to include in their negative dataset other ncRNAs that can fold into a hairpin structure similar to that of pre-miRNAs. MicroPred another SVM-based miRNA prediction tool published by Batuwita and Palade [20], used the same 23 features utilized by miPred but with improved training datasets. MicroPred training datasets included 691 experimentally verified human miRNA precursors as the positive dataset and 8494 pseudo hairpins and 754 other ncRNA that can fold into hairpin-like structures as the negative dataset. The improved training dataset of microPred resulted in an improvement in performance, where microPred scored a higher average of sensitivity and specificity (93.58%) in comparison withthat of miPred (91.01%). miRD is another SVM-based prediction tool published by Yuanwei *et al.,* [24]. miRD used two sets of features (one for multi-stem pre-miRNAs and one for single-stem pre-miRNAs) to construct two independent SVM models. A boosting method was then applied to combine these models. miRD is used to give the probability of a candidate pre-miRNA to be a real one, or used to predict probable pre-miRNAs from a set of sequences resulting from deep sequencing data. Table **2** compares the performance of these non-comparative tools based on sensitivity, specificity and accuracy.

All previously mentioned miRNA prediction methods require a well-annotated genome of the organism understudy, a large sample of experimentally verified miRNAs as the positive dataset and a large number of pseudo miRNAs as the negative dataset. Although these requirements might not be considered a problem when predicting miRNAs for a well-studied organism like humans and *Caenorhabditis elegans*, they become serious problems when studying other less-studied organisms. The importance of miRNAs in post-transcriptional regulation, the lack of a sufficient number of known miRNAs and the significant number of poorly annotated genomes collectively call for novel methods that can overcome these hurdles [25]. MiRank [25], a novel method, is capable of overcoming these problems by implementing a random walk machine-learning algorithm that does not require a large number of positive miRNAs, does not utilize data from annotated genomes and does not require any negative dataset. MiRank fragments the genome of study into smaller sequences, each capable of being an miRNA precursor and termed a putative miRNA. It then represents each putative miRNA and the miRNAs precursors of the positive set (query sample) as vertices on a weighted graph (G=V, E). Each vertex is represented by a set of 36 features that reflect unique characteristics of true miRNAs. An edge is introduced between each putative and real miRNA precursor if they are close to each other. The weight (W) of the edge will then quantify the relationship between these two vertices. Based on the similarity (closeness) of the putative miRNAs to the positive set, a relevancy value is given to each putative miRNA and the result is sorted in descending order with the most possible miRNA precursor candidates ranked at the top. To validate the performance of this tool, the authors applied their tool on the human genome, where it was separated into fragments of 90 nucleotides. Then, these fragments were tested for their ability to form hairpin secondary structures based on several fold ingthres holds. Finally, from the fragments that passed the folding test,1000 fragments were randomly selected and a number of true human pre-miRNAs were added to the pool of sequences. To assess the performance of their tool, they used two performance measures defined as follows:

Table 3. A general Comparison Between Comparative and Non-Comparative Tools

Tool Type	Homology	Training/Test Datasets	Recommended Use
Comparative	Uses homology and sequence alignment at some point during prediction process	Independent-weakly dependent	To discover highly conserved miRNAs in selected species
Non-Comparative	Doesn't consider homology and conservation during prediction	Strongly dependent	To discover species-specific miRNAs and novel miRNAs

$$\text{Recall} = TP/(TP + FN) \quad (1)$$

$$\text{Precision} = TP/(TP + FP) \quad (2)$$

where TP, FP, FN are numbers of true positive predictions, false positive predictions and false negative predictions, respectively. The authors calculated each of these measures for different numbers of query samples (positive dataset) including 1, 5, 10, 15, 20 and 50.The best results were obtained when they used 50 query samples, where miRank was able to distinguish 0.682 of the true pre-miRNAs from the pool of putative pre-miRNAs with a precision of 0.939. Predicting miRNA precursors is the first step in studying miRNAs. It is important after predicting a miRNA precursor to validate it experimentally and identify its gene targets and functions. It is also important to understand the impact of this miRNA on certain diseases and biological pathways.

Computational tools are also being developed to help scientists analyze novel miRNA precurs or sand discover their biological/medical implications [26]. Table **3** provides a general comparison between comparative and non-comparative tools.

MIRNA DEEP SEQUENCING IDENTIFICATION TOOLS

Deep sequencing has allowed for the identification of novel miRNAs with great sensitivity and led to sharp increase in their discovery rate. A key difficulty in deep sequencing lies in separating miRNAs from other RNA species during sample preparation. After sample preparation ligands are attached to both ends and cDNAs are produced by reverse transcription. Depending on the sequencing technology, millions of reads can be generated resulting in a need for extensive bioinformatics analysis. A typical bioinformatics analysis of deep-sequencing miRNAs generated from Next Generation Sequencing (NGS) platforms involves filtering out other small RNAs, mapping to a reference genome and/or mapping to miRNA databases for the identification of known miRNAs and prediction of novel miRNAs.

In an attempt to study the importance and function of miRNAs in peanut Wang *et al.,* performed deep sequencing of all miRNAs in peanut using high-throughput Solexa sequencing technology [27]. Their study led to the discovery of 14 novel and 22 conserved miRNA families from peanut, which was verified using qRT-PCR analysis. Their bioinformatics analysis involved filtering out all rRNA, tRNA, snRNA, and snoRNA, as well as reads containing the polyA tail and then comparing the remaining reads against rice and *Arabidopsis* ncRNAs deposited in the NCBI Genbank database and Rfam8.0 database. Then, the unique small RNA sequences were used to do a Blastn search against the miRNA database, miRBase 13.0, in order to identify conserved miRNAs in peanuts. The remaining of the small RNA sequences were used to perform Blastn searches against peanut ESTs in order to obtain precursor sequences for novel potential miRNAs. Only precursor matches that were capable of forming a perfect stem-loop structure in addition to adhering to other criteria were considered as novel peanut miRNAs. In their attempt to discover novel miRNAs expressed in peanuts they mostly relied on known peanut ESTs without using any machine-learning tool to predict miRNAs *de novo*.

Several tools were developed to aid in predicting and validating miRNAs produced from deep sequencing and data generated from NGS. miRDeep [28] one of the first tools developed for deep-sequencing dataemploys a probabilistic model of miRNA biogenesis to score compatibility of the position and frequency of sequenced RNA with the secondary structure of the miRNA precursor. miRDeep assigns a likelihood score that a predicted miRNA is a true mature miRNA and the authors used *C. elegans* data and data they generated by deep sequencing human and dog RNAs to validate the accuracy of the results. Using this tool they were able to predict ~230 novel miRNAs of which 4 *C. elegans*mi RNAs were validated by northern blot. A newer version of miR Deep termed miRDeep2 [29] was recently developed, which offers significant improvements in resources consumption and accuracy. miR analyzer [30] is a web-based tool that takes a file containing sequence reads and its respective copy numbers to perform the following; i). Identifies all known miRNAs annotated in the miRBase, ii). Finds all perfect matches against other libraries of transcribed sequences and iii). Predicts new miRNA with high accuracy using a machine learning approach based on random forests after filtering out sequences from the first two steps to reduce the numbers of false positives. DSAP [31] is a popular tool for analyzing deep-sequencing miRNA data generated by Solexa. DSAP doesn't require a target genome and instead clusters the reads into groups that are mapped against existing RNA/miRNA databases. DSAP is used for the identification of known miRNAs and RNAs from deep-sequencing and doesn't perform any sort of miRNA prediction thus it is not suitable for the discovery of novel miRNAs. Tool selection for deep sequencing miRNA data analysis can lead to significantly different results, a careful examination of available tools and their purposes is required before carrying on any bioinformatics analysis. *Vladimirov et al.,* [32] and *Shen et al.,* [33] each carried a performance evaluation of popular tools for deep sequencing miRNA data analysis using available and their own generated datasets. Computational time, accuracy, sensitivity and species-specific performance are amongst the most important criteria for tool selection.

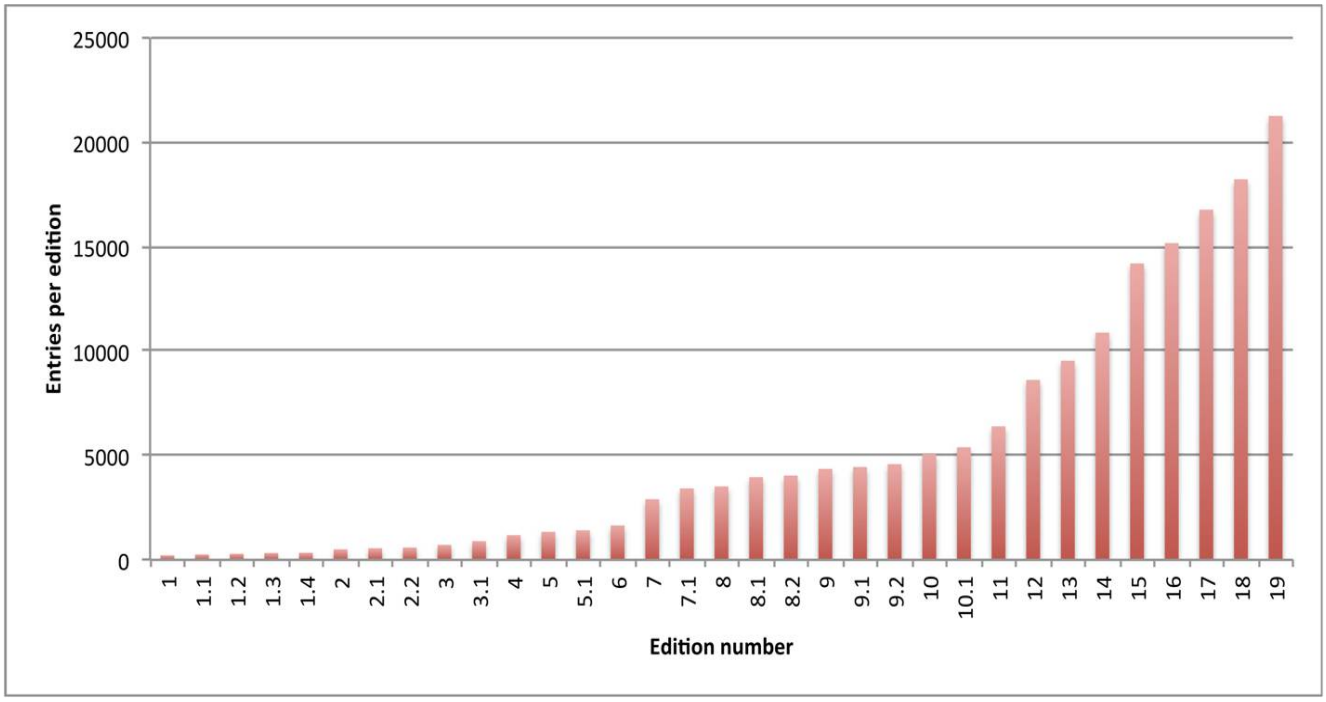

Fig. (3). The number of published miRNAs available at MiRBase is increasing very rapidly ever since the online repository started in 2002. (9)

Deep sequencing of miRNA not only can be used to discover novel miRNAs but can also be applied to quantify the expression-level of detected miRNAs. A recent study by Croce *et al.,* identified a nine-micro RNA signature that differentiated invasive from *in situ* carcinoma in breast cancer [34]. Martelli *et al.,* utilized deep sequencing of miRNAs on SOLiD platform to comprehensively and accurately profile the entire miRNA population expressed by endothelial cells exposed to hypoxia [35]. Their bioinformatics pipeline identified more than 400 annotated miRNAs/miRNAs* with miR-21 and miR-126 totaling almost 40% of all miRNA abundance. Their bioinformatics analysis and validation using qPCR resulted in the discovery of 18 high-confidence novel miRNAs, two of which were significantly down-modulated by hypoxia. In another study that utilized deep sequencing on SOLiD, *Guo et al.,* characterized the cellular microRNA profile involved in the development of congenital heart malformation, through the investigation of single ventricle (SV) defects [36]. They discovered 38 down-regulated and 10 up-regulated miRNAs in differentiated SV cardiac tissue, compared to control cardiac tissue.

To keep up with the vast amount of deep-sequencing miRNA data *Griffiths-Jones et al.,* mapped reads from short RNA deep-sequencing experiments to microRNAs in miRBase (the primary online repository for all microRNA sequences and annotation) and developed web interfaces to view these mappings [37].The user can view all read data associated with a given microRNA annotation, filter reads by experiment and count, and search for microRNAs by tissue- and stage-specific expression. These data can be used as a proxy for relative expression levels of microRNA sequences, provide detailed evidence for microRNA annotations and alternative isoforms of mature microRNAs, and allow us to revisit previous annotations.

COMPUTATIONAL RESOURCES BEYOND MIRNA PREDICTION

Computational resources for miRNA study go beyond just miRNA prediction and discovery. Many tools are being developed to help researchers better store/retrieve, validate, predict target sites and functionally study the newly discovered miRNAs.

MIRNA DATABASES

MiR Base 18.0 [9] and miRGen 3.0 [38] are two commonly used databases for miRNAs. The former is a repository where newly discovered micro RNAs are deposited with their genomic locations, sequences and references. The latter is a database that provides information on the genomic position of miRNAs (e.g., chromosome number and genes containing the miRNA sequence) and nearby features suchas transcription starting sites and transcription factor binding sites. MiRNA databases are witnesses to thegreat increase in the number of discovered miRNAs as a result of the use of computational tools in the field of miRNA study, where the number of submitted miRNAs to the miR Base has doubled manytimes since its first release (Fig. **3**).

Another type of miRNA database isthe miRNA-target interaction database. Such databases contain information about experimentally validated miRNA-target interactions. MiRecords [39] is a good example of such a database, where it lists for each stored miRNA all its validated targets and manually curated results under the Validated-Targets component. It also lists predicted targets based on 11 established miRNA target prediction tools under the Predicted-Targets component. The validated targets component of the database contains 2286 interactions between 548 miRNAs and 1579 target genes in nine species

(update 25 November 2010). miR2 Disease [40] is a manually curated database that aims to provide information regarding miRNA-related pathologies. TarBase 6.0 [41] is the most recent version of DIANA's Lab TarBase first released in 2005. The sixth version of TarBase aims at providing a significant increase of available miRNA targets derived from all contemporary experimental techniques (gene specific and high-throughput), while incorporating a powerful set of tools in a user-friendly interface. The authors developed a text mining-assisted literature curation tool in order to reduce the necessary time for manuscript curation and introduced a new relational database schema to accommodate present and future updates to the database. The new database includes 65,814 experimentally validated miRNA–gene interactions which is a 50-fold increase of entries from the latest TarBase version and a 16.5- to 175-fold increase from all the other available manually curated databases.

miRNA TARGET PREDICTION AND FUNCTION ANALYSIS

miRNA target prediction is a very important step towards understanding the regulatory function of thousands of recently discovered miRNAs. Experimental methods for miRNA target identification are often not feasible and difficult; thus computational methods for miRNA target prediction are expected to remain important for miRNA target studies and as a means for directing related wet-lab experiments.

There are several target prediction tools available online, where the user inputs the sequence ofthe miRNA understudy and the tool outputs a list of predicted targeted genes based on its computational algorithm. DIANAmicroT3.0 [42], miRanda-mirSVR [43], PITA [44] and others are among the most popular miRNA target prediction tools available online.

Finally, many researchers are interested in finding out whether miRNAs they have identified are associated with any disease or biological process. Online tools like DIANA-mirPath [45] are being used to address *in silico* miRNA function analysis. DIANA-mirPath takes all combinations of all of the predicted targets of the miRNA understudy and searches for enrichment against all known KEGG pathways. The authors argue that by knowing the biological pathways of the miRNA targets, the user can infer the functional importance of the miRNA.

CONCLUSION

MiRNAs are important post-transcriptional regulators that are involved in many cellular processes, such as differentiation, proliferation and apoptosis and are linked to many diseases such as on cogenesis. Given their biological importance, miRNAs are emerging therapeutic targets in a broad range of diseases and are expected to develop into a novel armada of more powerful and mechanism-oriented the rapeutics. Discovering novel miRNAs is needed to further our understanding of their biological functions and relation to disease, and given the limited abilities to discover them using classical wet-lab experimental methods there is a need for *in silico* miRNA prediction methods. *In silico* miRNA prediction benefited from the few number of already experimentally discovered miRNAs at the time, and relied on their profound conservation across species to discover novel ones, this gave rise to the Comparative miRNA prediction tools which rely on miRNA conservation for novel miRNA discovery. Comparative tools are unable to predict species-specific miRNAs (non-conserved miRNAs) and this led to the introduction of Non-Comparative miRNA prediction tools, which utilize machine-learning algorithms and miRNA specific features derived from experimentally verified miRNAs for novel miRNA prediction. With the advancement in sequencing technologies and wide availability of NGS, discovery of novel miRNAs via deep sequencing became the method of choice. Deep sequencing of miRNAs provides means of studying the complete profile of miRNAs at a certain condition or cell type and has resulted in steep increase in the discovery rate of novel miRNAs. *In Silico* miRNA prediction must be followed by wet-lab experimental validation of the top miRNA candidates to confirm their expression. MiRNA discovery is only the first step of miRNA study that must be followed by target and function analysis for a complete understanding of its biological importance. miRNAs will continue to garner significant attention from the scientific community and computational tools will be expected to continue to deliver better results.

LIST OF ABBREVIATIONS

miRNA	=	MicroRNA
ncRNA	=	non-coding RNA
pri-miRNA	=	preliminary MicroRNA
pre-miRNA	=	precursor MicroRNA
miRISC	=	miRNA-induced silencing complex
SVM	=	Support Vector Machine
UTR	=	untranslated region
HMM	=	Hidden Markov Model
CAGR	=	Cancer Associated Genomic Regions
NGS	=	Next Generation Sequencing
RNA	=	Ribonucleic acid

AUTHORS' CONTRIBUTIONS

TR developed the concept of the review. TM and TR wrote the review.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

T.R. and T.M. are supported by King Abdullah University of Science and Technology.

REFERENCES

[1] Anastasis O, Martin R, Panayiota P. MicroRNAs and Cancer-The Search Begins! IEEE Trans Inform Technol Biomed 2009; 13: 11.

[2] Li M, Li J, Ding X, He M, Cheng SY. microRNA and cancer. Aaps J 2010; 12: 309-17.

[3] Lee Y, Kim M, Han J, *et al*. MicroRNA genes are transcribed by RNA polymerase II. EMBO J 2004; 23: 4051-60.

[4] Starega-Roslan J, Koscianska E, Kozlowski P, Krzyzosiak WJ. The role of the precursor structure in the biogenesis of microRNA. Cell Mol Life Sci 2011; 68: 2859-71.

[5] Ruby JG, Jan CH, Bartel DP. Intronic microRNA precursors that bypass Drosha processing. Nature 2007; 448: 83-6.

[6] Pratt AJ, MacRae IJ. The RNA-induced silencing complex: a versatile gene-silencing machine. J Biol Chem 2009; 284: 17897-901.

[7] Miranda KC, Huynh T, Tay Y, *et al*. A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. Cell 2006; 126: 1203-17.

[8] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005; 120: 15-20.

[9] Griffiths-Jones S. The microRNA Registry. Nucleic Acids Res 2004; 32: 109-11.

[10] Berezikov E, Cuppen E, Plasterk RH. Approaches to microRNA discovery. Nat Genet 2006; 38 Suppl: S2-7.

[11] Hu J, Xu Y, Hao J, Wang S, Li C, Meng S. MiR-122 in hepatic function and liver diseases. Protein Cell 2012; 3: 364-71.

[12] Tan Gana NH, Victoriano AF, Okamoto T. Evaluation of online miRNA resources for biomedical applications. Genes Cells 2012; 17: 11-27.

[13] Mendes ND, Freitas AT, Sagot MF. Current tools for the identification of miRNA genes and their targets. Nucleic Acids Res 2009; 37: 2419-33.

[14] Zhang BH, Pan XP, Cox SB, Cobb GP, Anderson TA. Evidence that miRNAs are different from other RNAs. Cell Mol Life Sci 2006; 63: 246-54.

[15] Zeng Y, Cullen BR. Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. Nucleic Acids Res 2004; 32: 4776-85.

[16] Ng KL, Mishra SK. *De novo* SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics 2007; 23: 1321-30.

[17] Pasquinelli AE, Reinhart BJ, Slack F, *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature 2000; 408: 86-9.

[18] Lindow M, Gorodkin J. Principles and limitations of computational microRNA gene and target finding. DNA Cell Biol 2007; 26: 339-51.

[19] Clote P, Ferre F, Kranakis E, Krizanc D. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA 2005; 11: 578-91.

[20] Batuwita R, Palade V. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics 2009; 25: 989-95.

[21] Hertel J, Stadler PF. Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. Bioinformatics 2006; 22: e197-202.

[22] Szafranski K, Megraw M, Reczko M, Hatzigeorgiou AG. Support vector machines for predicting microRNA hairpins. *BIOCOMP* 2006; 270-6.

[23] Oulas A, Boutla A, Gkirtzou K, Reczko M, Kalantidis K, Poirazi P. Prediction of novel microRNA genes in cancer-associated genomic regions--a combined computational and experimental approach. Nucleic Acids Res 2009; 37: 3276-87.

[24] Yuanwei Z, Yifan Y, Huan Z, *et al*. Prediction of novel pre-micrornas with high accuracy through boosting and SVM. Bioinformatics 2011; 27(10): 1436-7.

[25] Xu Y, Zhou X, Zhang W. MicroRNA prediction with a novel ranking algorithm based on random walks. Bioinformatics 2008; 24: 50-8.

[26] Alexiou P, Maragkakis M, Hatzigeorgiou AG. Online resources for microRNA analysis. J Nucleic Acids Invest 2011; 2: e4.

[27] Chi X, Yang Q, Chen X, *et al.* Identification and characterization of microRNAs from peanut (Arachis hypogaea L.) by high-throughput sequencing. PLoS One 2011; 6: e27530.

[28] Friedlander MR, Chen W, Adamidi C, *et al.* Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol 2008; 26: 407-15.

[29] Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 2012; 40: 37-52.

[30] 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; 39: W132-8.

[31] Huang PJ, Liu YC, Lee CC, *et al*. DSAP: deep-sequencing small RNA analysis pipeline. Nucleic Acids Res 2010; 38: W385-91.

[32] Williamson V, Kim A, Xie B, McMichael GO, Gao Y, Vladimirov V. Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. Brief Bioinform 2012. doi: 10.1093/bib/bbs010

[33] Li Y, Zhang Z, Liu F, Vongsangnak W, Jing Q, Shen B. Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. Nucleic Acids Res 2012; 40: 4298-305.

[34] Volinia S, Galasso M, Sana ME, Wise TF, Palatini J, Huebner K, *et al.* Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. Proc Natl Acad Sci U S A 2012; 109: 3024-9.

[35] Voellenkle C, Rooij J, Guffanti A, *et al*. Deep-sequencing of endothelial cells exposed to hypoxia reveals the complexity of known and novel microRNAs. RNA 2012; 18: 472-84.

[36] Yu ZB, Han SP, Bai YF, Zhu C, Pan Y, Guo XR. MicroRNA expression profiling in fetal single ventricle malformation identified by deep sequencing. Int J Mol Med 2012; 29: 53-60.

[37] Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 2011; 39: D152-7.

[38] Megraw M, Sethupathy P, Corda B, Hatzigeorgiou AG. miRGen: a database for the study of animal microRNA genomic organization and function. Nucleic Acids Res 2007; 35: D149-D55.

[39] Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 2009; 37: D105-10.

[40] Jiang Q, Wang Y, Hao Y, *et al*. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 2009; 37: D98-104.

[41] Vergoulis T, Vlachos IS, Alexiou P, *et al.* TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Res 2012; 40: D222-9.

[42] Maragkakis M, Reczko M, Simossis VA, *et al*. DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res 2009; 37: W273-W6.

[43] Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 2010;11: R90.

[44] Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet 2007; 39: 1278-84.

[45] Papadopoulos GL, Alexiou P, Maragkakis M, Reczko M, Hatzigeorgiou AG. DIANA-mirPath: Integrating human and mouse microRNAs in pathways. Bioinformatics 2009; 25: 1991-3.

Received: June 18, 2012 Revised: August 4, 2012 Accepted: August 16, 2012