STUDY OF microRNA-mRNA INTERACTIONS RELATED WITH CANCER

KARATHANASIS NESTORAS

Ph.D. 2013

Study of miRNA – mRNA Interactions Related with Cancer

A Doctoral Dissertation

by

Karathanasis Nestoras

Department of Biology, University of Crete

and

Institute of Molecular Biology and Biotechnology, FORTH

2013

This thesis is dedicated to

my family.

TABLE OF CONTENTS

LIST OF FIGURES

4.8. miRNA-sensor assay using luciferase expression as an indicator of miRNA activity after transfection of heLa cells with various constructs.

LIST OF TABLES

ABSTRACT

MicroRNAs belong to the large family of small non coding RNAs. They regulate protein synthesis by binding to their mRNA targets causing mRNA degradation or translational repression. A large number of miRNAs have been associated with cancer because they are often found to be located within cancer associated genomic region (CAGRs/FRA) to target cancer-related genes, and to be differentially expressed in tumor compared to normal tissues. Previous work in the Computational Biology lab had identified four new putative miRNA genes that were located within CAGR. However their mature molecules and their association with cancer phenotypes were unknown. My thesis focuses on resolving these two issues, using a combination of theoretical and experimental techniques. The specific aims of this work are:

- The development of a mature miRNA prediction algorithm(Chapter II, III)
- The identification of the mature miRNA molecules of the newly identified miRNA genes via a combination of computational and experimental methods (Chapter IV)
- The utilization of a target prediction algorithm to predict and experimentally verify interactions between the mature molecules and cancer-related genes Chapter IV).

ΠΕΡΙΛΗΨΗ

Τα microRNA είναι μικρά μη κωδικοποιά μόρια RNA τα οποία προσδένονται στην 3’ αμετάφραστη περιοχή (3’UTR) του mRNA στόχου και οδηγούν σε καταστολή της μετάφρασης ή/και την αποικοδόμηση του. Έχουν συνδεθεί με διάφορα είδη καρκίνου, μέσω της εμφάνισής τους σε γενωμικές περιοχές που σχετίζονται με καρκίνο (CAGR/FRA), επειδή στοχεύουν γονίδια που εμπλέκονται σε καρκίνο ή επειδή η έκφραση τους εμφανίζεται διαφοροποιημένη σε καρκινικούς ιστούς. Το εργαστήριο της Δρ. Ποϊράζη ανακάλυψε πρόσφατα τέσσερα καινούργια πρόδρομα microRNA σε CAGR, χωρίς ωστόσο να είναι γνωστά τα ώριμα μόρια και η ακριβής σχέση τους με τον καρκίνο. Στόχοι τη παρούσας διατριβής είναι:

- Η δημιουργία ενός υπολογιστικού εργαλείου για την πρόβλεψη των ώριμων μορίων των miRNA, (περιγράφεται στο κεφάλαιο II και III).
- Η πειραματική εύρεση των ώριμων μορίων που παράγονται από τέσσερα πρόδρομα miRNAs, (περιγράφεται στο κεφάλαιο IV).
- Η υπολογιστική πρόβλεψη και πειραματική επιβεβαίωση αλληλεπιδράσεων μεταξύ των ώριμων μορίων και γονιδίων που έχουν συσχετιστεί με τον καρκίνο. (περιγράφεται στο κεφάλαιο IV).

ACKNOWLEGMENTS

First I would like to acknowledge Dr. Poirazi, Dr. Kalantidis, Dr. Kardasis and Dr. Tsamardinos for their valuable guidance and supervision. In addition the member of Dr. Kalantidis' lab, specifically Elena Dadami and Anastasis Oulas for helping me in Northern blotting and in miRNA target prediction, respectively. Also the member of Dr. Poizari's lab, George Kastellakis for his help in developing the web interface of MiRduplexSVM and Pavlos Pavlidis for the valuable conversations on the statistical analysis of the results. I would like to thank Dr. Tsamardinos lab-member Angelos Armen for his contributions regarding the initial development of aspects of MiRduplexSVM method. Finally I would like to mention that this research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund».

CHAPTER I - INTRODUCTION

BIOGENESIS OF MICRO-RNAS

MicroRNAs are small (~22 nts), single-stranded non-coding RNAs with a key regulatory role in both animals and plants. The primary transcripts of microRNA genes (pri-miRNAs) consist of a stem-loop (hairpin) structure extended with long single-stranded tails. The tails are detached (in animals) by the Microprocessor complex, whose core component is the RNase III enzyme Drosha, leaving a hairpin-shaped, ~60-70 nts long intermediate with a characteristic 3' overhang of ~2 nt, the miRNA precursor (pre-miRNA).

Two models have been proposed on how a pri-miRNA is processed to produce a pre-miRNA. According to the first model, Drosha or the holoenzyme with Drosha providing the catalytic activity, selects an RNA hairpin bearing a terminal loop that is no less than 10 nucleotides long, and cuts ~22 nucleotides from the terminal loop – stem junction to produce a pre-miRNA, Figure 1.1 [6]. According to the second model, the cleavage site is determined mainly by the distance (~11 base pairs) from the stem – single stranded tails junction, Figure 1.2 [3]. Note that some pre-miRNAs, the so-called mirtrons, have a similar structure with regular pre-miRNAs but enter the miRNA pathway without undergoing processing by Drosha, i.e. without undergoing the pri-miRNA stage [8]. Irrespectively of its production process, the pre-miRNA is then exported to the cytoplasm, where it is processed by another RNase III termed Dicer. Dicer

cleaves the pre-miRNA at a certain distance (~22 nt) from the overhang created by the Microprocessor[9], leaving an RNA duplex with 3' overhangs of ~2 nts called the miRNA-miRNA* duplex. For each individual duplex, one of its strands, the mature miRNA, is loaded into a RISC (RNA-induced Silencing Complex), where it performs its regulatory functions. The other strand, the miRNA*, is degraded. It may also be the case that both strands of the duplex correspond to a mature miRNA; however, only one strand becomes functional each time, but with similar frequency [10].

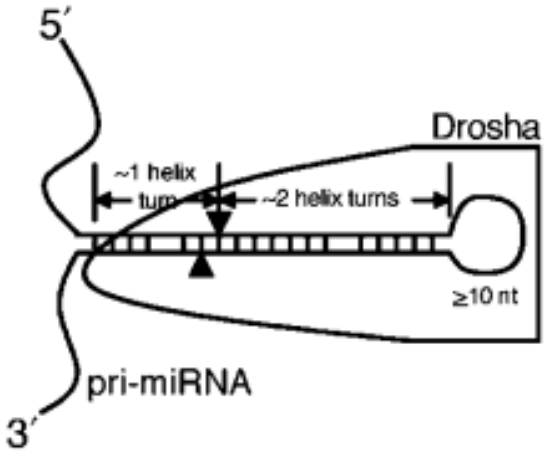

Figure 1.1. A model of how a pri-miRNA is processed to produce a pre-miRNA. In this model, Drosha, or a holoenzyme with Drosha providing the catalytic activity, selects an RNA hairpin bearing a terminal loop that isX10 nt long, and cutsB22 nt, orB2 helix turns, from the terminal loop/stem junction to produce a premiRNA. Efficient processing, and possibly recognition, also requires an extended (B10 bp), mostly double-stranded region located beyond the pre-miRNA stem. Figure adopted [6]

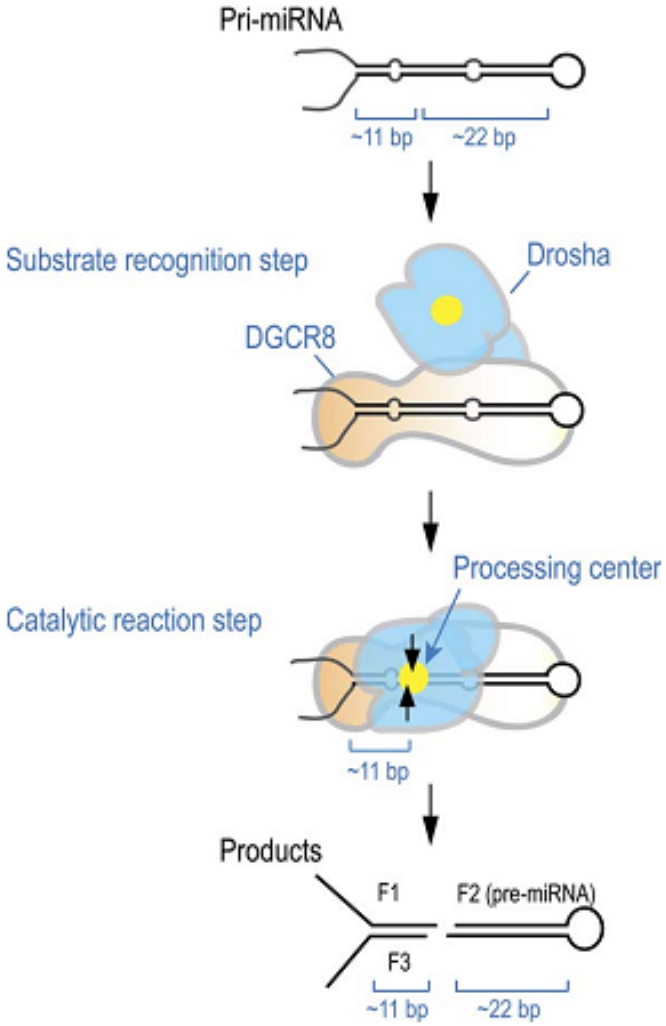

Figure 1.2. A "ssRNA-dsRNA Junction Anchoring" Model for the Processing of pri-miRNA. DGCR8 may play a major role in substrate recognition by directly anchoring at the ssRNA-dsRNA junction. DGCR8 also interacts with the stem of 33 bp and the terminal loop for a full activity although the terminal loop structure is not critical for DGCR8 binding and cleavage reaction. After the initial recognition step, Drosha may transiently interact with the substrate for catalysis. The processing center (yellow circle) of Drosha is placed at11 bp from the basal segments. Figure adopted from [3]

FUNCTION OF MICRO-RNAS

Micro-RNAs are controlling protein synthesis by binding to their mRNA targets through sequence complementarity rules. This binding leads to mRNA degradation or translational repression [10]. First, the mature miRNA is loaded into a RISC forming a miRISC complex. This complex, whereby the miRNA provides the specificity on target identification, regulates mRNAs expression. At no-repression conditions, mRNAs recruit initiation factors and ribosomal subunits and form circularized structures that enhance translation. On the contrary, when miRISCs bind to mRNAs, they inhibit translation through various mechanisms. They can repress translation initiation at the cap recognition stage, or during the recruitment of the 60s ribosome's subunit. Instead, they can induce deadenylation of the mRNA and thus inhibit circularization of the mRNA. Also, they can induce ribosomes to drop off prematurely repressing a postinitiation stage of translation. Finally, they can promote mRNA degradation by inducing deadenylation followed by decapping[11].

MICRO-RNAS AND CANCER

MicroRNAs play a fundamental role in many of the cell functions, as proliferation and differentiation [12], while they have been also found to relate with a number of diseases [13, 14], including many types of cancer [15]. Indicative examples are microRNAs let-7 [16], mir-15a/mir-16-1 family [17] and the neighboring mir-143 and mir-145 [7], whose expression is limited in specific types of cancer

implying a possible tumor suppressor role. On the contrary, mir-17-92 family [18-20] and mir-155/BIC [21] are overexpressed in several types of cancer and they have been associated to oncogenic procedures. In addition, a significant number of microRNAs is located on genomic regions which are susceptible to genetic alterations as deletions, duplications, and single mutations. These regions are known as Cancer-Associated Genomic Regions or CAGRs, and Fragile Sites or FRA[4]. MiRNA genes located within, or in close proximity, to these regions have been suggested to be associated with chromosomal events leading to carcinogenesis, as graphically illustrated in Figure 1.3.

Recent findings showed that mir-15a and mir-16a are located within the region 13q14, which is deleted in more than half cases of B cell chronic lymphocytic leukemias (B-CLL) patients[22]. Detailed analysis showed that mir-15a and mir-16a are actually located within a 30-kb distance from the region which is deleted and total or partial decrease of both genes' expression is observed in 68% of B-CLL cases [1]. Deletions in region 13q14 are also observed in prostate cancer with a 60% incidence, in multiple myelomas (16%-40%) as well as in other types of cancer, indicating that one or more tumor suppressor genes which are located on this region are implicated in the pathogenesis of human tumors [22]. Finally, it is known that approximately 30% of microRNA genes are located within introns of other genes and it is very likely that they are transcribed along with their host genes [23]. Consequently, in cases of genes which are overexpressed in some types of cancer, a thorough examination of their introns could lead to the discovery of new microRNAs with a crucial role in carcinogenesis.

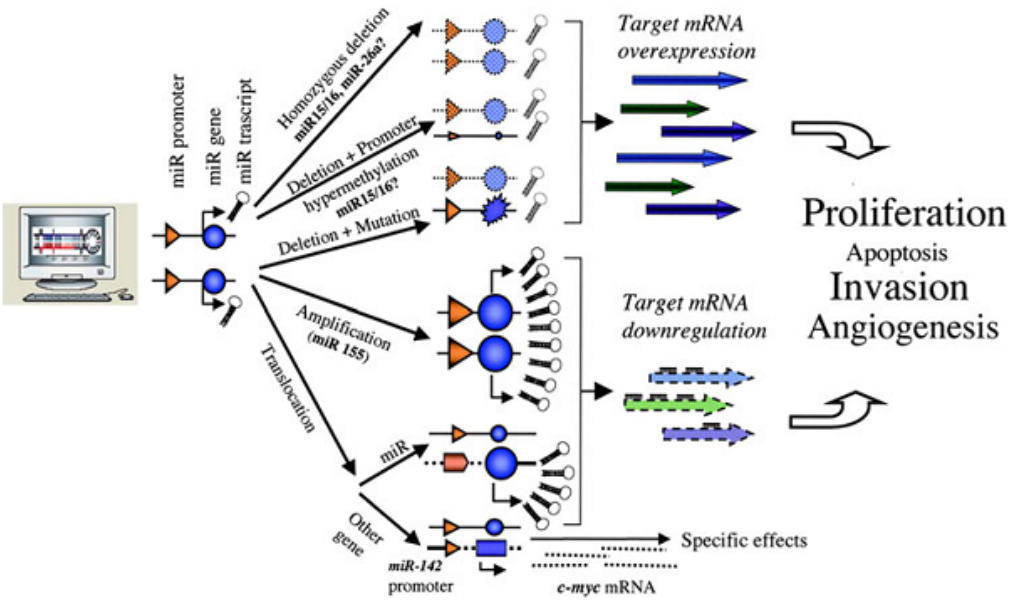

Figure 1.3. MiRNAs as cancer players. Computational prediction initiates the search for putative miRNAs that play a role in tumorigenesis. Some of these proposed mechanisms are experimentally proven, like the deletion ofmiR-15a/miR-16a cluster in B-CLL [1, 2], the c-myc over-expression by the reposition near a putative miR promoter [4], or miR143/miR-145cluster down regulation in colon cancers[7]. Figure adopted with permission from Callin et al. [4].

MICRO-RNA-RELATED COMPUTATIONAL TOOLS

Micro-RNAs have been associated with more than 150 diseases from several kinds of cancers to cardiovascular diseases, cerebellar neurodegeneration e.t.c. [24]. In addition high throughput sequencing experiments are performed to discover new miRNA mature molecules which could play a key role in disease progression. These miRNAs are often used as biomarkers and their biological significance is investigated through the prediction and subsequent experimental verification of miRNA-mRNA interactions. Although sequencing technologies evolve rapidly, they exhibit significant limitations. They are cost and time

consuming and usually only miRNAs which are in high levels are being sequenced. Furthermore the predictions of miRNA-mRNA interactions it is not a trivial task which can be perform manually. For these reasons several computational tools have been developed to complement experimental limitations and guide future experiments. These tools focus on different aspects of miRNA study; like miRNA gene/ mature prediction and miRNA-mRNA interaction predictions. In the following paragraph an outline of several tools is given.

miRNA gene prediction tools

The large amount of unexplored non-coding regions in the human genome combined with the increasing importance of miRNAs in multiple diseases highlights the need for fast, flexible, and reliable miRNA identification methods. Toward this goal a number of different computational methods have been used to identify miRNA genes. Early studies focused on scanning for hairpin structures conserved between closely related species such as Caenorhabditis elegans and Caenorhabditis briggsae [25, 26], or using homology between known miRNAs and other regions in aligned genomes like human and mouse[27]. Other approaches relied on conserved regions of synteny —conserved clustering of miRNAs in closely related genomes—to predict novel miRNAs [27]. Subsequent computational studies utilized profile-based detection [28] as well as secondary

structure alignment [29] of miRNAs using sequence conservation across multiple, highly divergent, organisms (i.e., mouse and fugu).

The main drawback of the abovementioned tools is that they undertake a pipeline approach by applying stringent cut-offs and eliminating candidate miRNAs as the pipeline proceeds [25, 26]. This results in the loss of numerous true miRNAs along the line. The use of homology by some tools [27-29] to detect novel miRNAs based on their similarity to previously identified miRNAs is another drawback. These methods obviously fall short when scanning distantly related sequences or when novel miRNAs lack detectable homologs.

The next generation of computational tools relied on more sophisticated machine learning algorithms such as support vector machines (SVMs) capable of taking into account multiple biological features such as free energy of the hairpin structure, paired bases, loop length, and stem conservation to predict novel miRNAs [30-37]. Two very effective computational studies utilized Hidden Markov Models (HMM) and a Bayesian classifier [38, 39], to simultaneously consider sequence and structure features at the nucleotide level for predicting miRNA genes. These studies, however, did not integrate conservation information in their algorithms, an important feature of the majority of miRNA genes. More recently two computational tools miRRim [40] and SSCprofiler [5] also employing HMMs proved to be very effective, achieving high performance on identifying miRNAs in the human genome.

With the advent of large scale, high throughput methods such as tiling arrays or deep sequencing the identification of novel miRNA genes is taking a different turn [41-43]. These methods are exceptionally useful as they produce large

datasets that offer a relatively accurate expression map for small RNAs in the genome. However, since large-scale expression data are usually limited by the specific tissue and developmental stage of their samples, only the coupling of such data to computational tools (as done in two recent studies [5, 40]) can facilitate rapid and precise detection of novel miRNAs, while at the same time giving greater credence to computational predictions.

miRNA mature prediction tools

Given the importance of miRNAs in gene regulation, several computational approaches have been developed to complement experimental ones. Most of them focus on the discovery of novel miRNA genes or possible mRNA targets of known miRNAs[44-49]. As part of miRNA gene discovery, these tools predict certain features of miRNAs such as the starting position of the mature miRNA [39, 50, 51], the Drosha cleavage site [52] (which coincides with the start of the mature miRNA on a pri-miRNA). Most computational approaches labeled as miRNA predictors are actually pre-miRNA predictors, in the sense that they identify candidate genomic regions that may form pre-miRNAs but rarely attempt to determine the position of the miRNA itself within them.

To the best of our knowledge, only six mature miRNA predictors have been proposed to date which use a machine learning approach.

I. *ProMir* [53] identifies human pre-miRNAs and their mature miRNAs by combining sequence and structural features in a paired hidden Markov model.

II. *MiRmat* [54], which is composed of two parts: the prediction of Drosha processing site and the identification of Dicer processing site. The free energy distribution pattern of the downstream part of pri-microRNA secondary structure and Random Forest algorithm are utilized to predict the mature miRNA sequence.

III. *MatureBayes* [55] was the first tool specifically developed to address the problem of mature miRNA identification. It utilizes a Naive Bayes classifier to identify 22-nt mature miRNA candidates based on sequence and secondary structure information of their miRNA precursors [55]. It generates one prediction per strand and uses the 2nts overhang rule to define the predicted miRNA* on the opposite strand, creating two hypothetical duplexes.

IV. *MiRPara* [56] is an SVM-based tool. It generates several independent predictions for each strand of any given hairpin. The size of the possible mature miRNAs ranges.

V. *MaturePred* [57] employs an SVM classifier to predict the region which is most likely to contain the mature miRNA molecule in each strand of a hairpin based on miRNA–miRNAs* features. It consists of two models, one specialized in plants and one specialized in mammals.

VI. Finally, *MiRdup* [58] is the latest tool that tackles the problem of mature miRNA identification. It does so by finding the most likely miRNA location within a given pre-miRNA. MiRdup is based on a random forest classifier

trained with experimentally validated miRNAs from miRbase, with features that characterize the miRNA–miRNA* duplex. MiRdup predicts the most probable miRNA duplex on a given hairpin.

In all cases, the results are amenable to improvement as performance accuracy with respect to the identification of the exact mature miRNA molecule remains low.

miRNA target prediction tools

The mode of action of the mature miRNA in mammalian systems is dependent on complementary base pairing primarily to the 3'-UTR region of the target mRNA, thereafter causing the inhibition of translation and/or the degradation of the mRNA. Searching through all human genes (~25,000) and/or other species for novel miRNA gene targets is a complicated task for which fast, flexible and reliable identification methods are required. Currently available experimental approaches working towards this goal are complex and sub-optimal [59]. Inefficiencies result from various sources, including difficulty in isolating certain miRNAs by cloning due to low expression, stability, tissue specificity and technical difficulties of the cloning and repression assay procedures, while selecting the right 3'UTR to investigate is often a challenging task of its own. Computational prediction of miRNA gene targets from 3'UTR genomic sequences is an alternative technique which offers a much faster, cheaper and effective way of identifying putative miRNA gene targets. Moreover, by predicting the location

of a miRNA gene target, these methods enable experimental biologists to concentrate their efforts on genomic regions more likely to contain novel genes that undergo miRNA regulation, thus facilitating the discovery process.

Due to the lack of negative data in this specific biological problem, the performance of current miRNA target prediction tools is largely dependent on the overall number of predicted targets. Some tools are very efficient in predicting true target sites (high sensitivity), but at the same time display an extremely large number of overall predictions (low specificity) [60-63] . In contrast, other tools display an overall high specificity and a relatively low sensitivity [64-66]. In order to provide an estimation of a false positive rate, false or mock miRNAs are often generated by randomly shuffling the nucleotide sequence of experimentally supported miRNAs [67]. Performing target prediction with these "mock" miRNAs can provide an estimation of the overall false positive rate of a miRNA target prediction tool.

Accurate prediction of novel miRNA gene targets requires the consideration of certain characteristic properties of the miRNA::target-mRNA interaction. These properties are based on either experimental [68-71], or computational evidence [26, 31-33, 39] and can be used to build a classification scheme or predictive model. For example, the foremost nucleotides at the 5'region of a mature miRNA sequence are considered crucial for recognizing and binding to the target mRNA. Initial research performed by *Kiriakidou et al*, *2004* [72] has shown that almost consecutive complementarity of the first 9 miRNA nucleotides to the 3'UTR of protein coding genes is a prerequisite for translational repression. Moreover, *Lewis et al, 2005* [64] showed that complementary motifs to nucleotides 2-7 of

miRNA (commonly referred to as the miRNA seed region) remain preferentially conserved in several species in a statistically significant manner [73, 74].

In general, it is believed that binding of at least 7 consecutive Watson-Crick (WC) base pairing nucleotides between the foremost 5'region of the miRNA and the mRNA target is required for sufficient repression of protein production [64, 72].

Based on the above mentioned evidence, miRNA target prediction programs rely heavily on sequence complementarity of the miRNA seed region (nucleotides 2-7) to the 3'UTR sequences of candidate target genes for identifying putative miRNA binding sites [65, 75]. Furthermore, most prediction tools make use of thermodynamics and evolutionary conservation at the binding site, in order to minimize false positives (increase specificity) [65, 75, 76]. Some tools utilize additional features such as, binding site structural accessibility [61, 77, 78], nucleotide composition flanking the binding sites [79, 80] or proximity of one binding site to another within the same 3' UTR [79, 81].

In summary, the general features employed for miRNA target prediction are: (i) sequence complementarily at the 5'region of the mature miRNA, better known as the seed region and commonly characterized by nucleotides 2-7, (ii) secondary structure of the miRNA::target-mRNA hybrid molecule and the overall thermodynamics of the interaction expressed in free energy (ΔG) and (iii) species conservation observed via the use of full genome sequence alignments.

In addition to computational tools, large scale, high throughput transcriptomic and proteomic methods, such as microarrays and pSILAC, have recently been used, often in conjunction with computational tools, for the identification of

novel miRNA gene targets [63, 82]. These methods are particularly useful as they can provide accurate protein repression data or gene expression data that may be correlated or anti-correlated to miRNA expression. Moreover, if such data is coupled to computational tools, it can facilitate rapid and precise detection of novel miRNA gene targets, while at the same time giving greater credence to computational predictions. One must keep in mind that such proteomic data may only provide in-direct evidence for target genes, as the signal obtained may be due to downstream effects and not a consequence of a direct interaction between miRNA and predicted target gene.

Next Generation Sequencing (NGS) methods have also been used for the prediction of miRNA genes, their mature sequences [41] and their downstream targets. One drawback of these techniques is that multiple small RNA sequences are often missed due to technical difficulties of the sequencing methodology, such as library construction. Moreover, not all small RNA and under expressed mRNA sequences detected by NGS methods are true miRNAs and targets respectively, unless, some physical interaction between the two sequences can be attributed to them. Experimental verification of miRNA targets can be achieved via the use of luciferase assays whereby the miRNA is expressed *in vitro*, while simultaneously expressing and monitoring the target messenger RNA linked to a luciferase reporter gene [2, 83, 84]. This assay provides an experimental verification of a direct interaction between the mature miRNA and the target gene and furthermore provides evidence that regulation is mediated via the miRNA silencing pathway. However, the extent to which this interaction

takes place in the intact system *in vivo* cannot be inferred from luciferase assays alone.

Currently, miRNA target prediction tools are freely available and are commonly built on sophisticated algorithms (i.e. machine learning) trained to recognize certain biological features of miRNA::target-mRNA interactions. Validation of computational methodologies is achieved using protein repression information from large scale proteomic studies (i.e. pSILAC) [82] as well as experimentally verified miRNA gene targets form online databases, like Tarbase (v5) [85]. Results from these analyses are used to obtain a comparison and prediction accuracy for existing target prediction tools. It is rare that target prediction tools are assessed for their ability to identify *de novo* biologically significant interactions. This can be achieved by directing predictions and high-throughput experiments towards answering a specific biological questions [86]. Such approaches can provide raw material for experimental biologists to investigate interesting biological questions, such as the molecular basis of a disease like cancer.

MOTIVATION

Based on the above indications, members of Dr.Poirazi's and Dr. Kalantidis laboratories had recently discovered new pri- or pre-microRNAs which are located on CAGRs [4, 5] via a combination of bioinformatics and experimental methods. More precisely, using SSCprofiler [5], an efficient miRNA gene prediction tool, four new pri- or pre-microRNAs were detected (Table 1) within regions which are deleted in prostate, colorectal and brain cancer (astrocytomas) [5], although the exact function and the potential association of these microRNAs with the respective cancer types were not investigated.

TABLE 1.1. FOUR NEW miRNAs

Candidate	Candidate Information	CAGR	Type of Cancer
1	chr9:123327358-123327460 st-	chr9:121153509-128793509	bladder
2	chr5:148958951-148959053 st-	chr5:144121683-156051683	prostate aggressiveness
2	chr5:148958951-148959053 st-	chr5:148181683-151101683	myelodysplastic syndrome
3	chr22:40863894-40863996 st+	chr22:31530000-43583971	colorectal
3	chr22:40863894-40863996 st+	chr22:31530000-42193557	astrocytomas
4	chr5:149984684-149984786 st-	chr5:144121683-156051683	prostate aggressiveness
4	chr5:149984684-149984786 st-	chr5:148181683-151101683	myelodysplastic syndrome

Exact location of the new pri- or pre-microRNAs on the genome, types of cancer related to alterations on these regions. (Compiled from [5])

SCOPE

The aim of the present thesis is twofold: (a) the development and application of computational methods for mature miRNA identification and miRNA target prediction and (b) the use of combined computational and experimental techniques to characterize the role of novel microRNA genes located on cancer related regions of chromosomes 9, 5 and 22. In more detail, this thesis focuses on:

- The generation of one computational tool for the prediction of mature molecules given a miRNA precursor.
- The determination of the mature molecule of the microRNA genes which are located on chromosome 9, 5 and 22.
- The refinement and application of a computational tool for the computational prediction and experimental evaluation of selected interactions between predicted microRNAs and cancer-related target-genes.

CHAPTER II – MiRduplexSVM

In this part of the thesis, we introduce the problem of identifying the miRNA:miRNA* duplex as a first step in predicting the mature miRNA. We adopt this approach because (a) the duplex is a necessary stage of miRNA biogenesis and may contain features that determine the functional molecule and (b) given the duplex, it is relatively easy to experimentally determine whether both, or which of the two duplex strands results in the mature miRNA(s).

We present a methodology that uses an appropriate representation of biological features combined with extensive optimization and training of SVM classifiers in order to generate predictive models of the miRNA:miRNA* duplex position on a hairpin sequence. Resulting models significantly outperform four existing tools, namely MatureBayes[55], MiRPara[56], MaturePred[57], and MiRDup[58] as well as a Simple Geometric Locator, a trivial method employing the position as the only predictor used for a baseline comparison. Moreover, our methodology can accurately predict the miRNA* given a known miRNA molecule and can be used to investigate the effects of mutagenesis on Drosha processing, leading to several experimentally testable predictions. In silico mutagenesis experiments performed on 142 hairpins suggest that both the distance from the single stranded stem junction and the distance from the terminal loop determine the Microprocessor's cleavage site with the latter playing a major role.

Several factors contribute to the success of the methodology: the definition of the problem (predicting the whole duplex vs. a single strand or end), the representation of the sequence with a fixed-length vector using zero padding in the middle, the inclusion of the duplex flanking sequences, the production of positive and negative training examples based on biological constraints and not the simple 2nt overhang rule, the optimization of the SVM hyper-parameters while avoiding overfitting, and the use of two cost hyper-parameters to address the problem of positive and negative training examples' imbalance.

METHODOLOGY

The key idea of the proposed methodology is to train and employ a full polynomial SVM model to score each possible duplex position on a hairpin sequence and select the highest scoring one as the final predicted location. The various steps of our methodology are presented in the following paragraphs.

Candidate Duplex Production

The production of all possible duplexes on a hairpin structure is employed to generate training examples for the SVM during the training phase but also to produce all duplexes to be scored at prediction time; the highest scoring one is the final prediction.

In a hairpin sequence, the counting of nucleotide positions starts from the 5' end and continues to the 3' end. A hairpin consists of a double-stranded part, the stem, and a sequence of unmatched nucleotides that connects the strands of the stem, called the terminal loop. The strand before the terminal loop is called the 5' arm of the hairpin while the other is called the 3' arm. The arms are not perfectly complementary but they form small loops and bulges.

A miRNA:miRNA* duplex consists of two hairpin substrings on each of the two arms, called the *5' strand* and the *3' strand* of the duplex. We can define a duplex by the positions of its four ends on the respective strands of the generating

hairpin sequence; we name them *k55*, *k53*, *k35* and *k33*. Figure 2.1A shows an example of a real hairpin (hsa-mir-17) with all the above quantities annotated. Notice that, because of the way of counting positions, *k55 < k53 < k35 < k33* holds.

Not all possible substrings on the two strands define a possible duplex. Several constraints that are obeyed by Nature (as far as we know) need to be satisfied: (1) Two strands that share no matching bases do not form a possible miRNA:miRNA* duplex. (2) The length of each duplex strand should lie within a certain range, which can be deduced from known miRNAs. (3) The duplex overhangs should also lie within specific ranges, which can be calculated using the training examples. The procedure for calculating the length of the overhangs is described for the k55 end and is similar for the k33 end. Based on the secondary structure of the hairpin, we identify the position of the base that matches k55 on the opposite strand, say k'55. The overhang length is obviously k33 – k'55. However, k55 does not always have a matching base. In this case, we move x positions to the left (inside the hairpin) until we find a base on the 5' arm that has a matching base on the 3' arm, say on position k'55. The overhang length is then computed as k33 – k'55 + x. Of course, when k55 has a matching base x=0 and the two computations coincide. (4) *k55 < k53 < k35 < k33* and (5) *k55, k53* and *k35, k33* need to be before and after the tip of the loop, respectively. To calculate the tip we identify the last matching nucleotides before the tip, which correspond to the loop start and loop end position, respectively. If the tip is T and the last matching nucleotides are X and X', then X < T < X' and T = X + ceil ((X'-X) / 2), ceil refers to rounding toward positive infinity. (6) Finally, the

distances of k55, k53, k35 and k33 from the loop tip should lie within specific ranges, which can be calculated using the training examples. Specifically the distance of k55 and k53 from the loop tip ranges between the minimum and the maximum distance observed in the training examples. The distance of the position k35 is equal or bigger than the minimum distance observed in the training examples. On the other hand the distance of k33 from the loop tip is equal or less than the maximum distance observed in the training set.

To assemble and detail the whole procedure together, we first predict the secondary structure of each hairpin using the RNAfold program [87] with the default parameters (-p -d0 -noLP -noPS). We then calculate the statistical distributions of the overhangs' and matures' lengths and of the distances of the four corners of the duplex from the loop tip, according to the secondary structures of hairpins in the given training set and remove the overhangs' outlier values (values that are above or below three times the standard deviation from the mean value). This is necessary to reduce the number of candidate duplexes, stemming from values that are too extreme and uncommon. We then produce all duplex sequences that correspond to each combination of values *k55, k53, k35, k33* that obey the constraints defined above. When only the first five constrains were used, the process of candidate production was named "All", whereas when all six constrains were used it was named "Selected". Contrary to earlier work [56], the k53 end can be positioned as far as the loop tip, allowing the identification of mature miRNAs that extent into the terminal loop.

These two methodologies result in the generation of ~10.000 candidate duplexes per hairpin, only one of which is the true duplex. During training, true duplexes

are labeled positive and the rest form the negative examples. During testing, the true duplex is occasionally not produced due to the restrictions on the possible ranges of the overhangs described above. In the experiments reported here, loss of true duplexes due to this filtering never exceeded 4%.

Duplex Vector Representation

Extensive experimentation was first performed in order to find the minimum set of features like sequence, structure or thermodynamics needed to obtain maximum accuracy. Here we describe the representation of each of these features in a vector.

Sequence

MiRNA:miRNA* duplexes used as input to the SVM are represented by a fixed-length numerical vector that contains nucleotide sequence information. Nucleotide sequences are converted to binary vectors, using a 1-of-4 encoding at each position: bases A, T, G and U are represented with four binary variables as 1000, 0100, 0010 and 0001, respectively. A fixed-vector representation becomes problematic when strand sequences are of variable size. One solution is to identify the maximum possible strand length and pad with zeros or some other special value at the end for the missing nucleotides. In this case, the suffix of the nucleotide sequences will be represented with different variables each time. However, it has been demonstrated that end structure and nucleotide sequence

are the primary determinants of Dicer specificity when processing double-stranded or short hairpin RNA [9]; we thus preferred a representation where it is the ends of the sequences that always correspond to the same variables. To do so we pad with zeros in the middle of a sequence, so that the first and the last nucleotides are always represented with the first and last variables, respectively. Zero padding is common in signal processing and while there may be better ways to treat missing information, it does not affect the estimation of model performance or invalidates any results.

As the flanking regions around Drosha and Dicer cut sites are critical for the identification of these sites [52], [55], we include the flanking regions at both ends of each duplex strand in the representation of a candidate duplex. If a flanking region extends beyond the arm's boundaries, zero padding at the beginning (for 5' end flanking regions) or at the end (for 3' end flanking regions) of the sequence takes place. The complete representation of a candidate duplex consists of the encoded nucleotide sequences of both its strands and their flanking regions. A preliminary version of this methodology is described in [88] and an example of the miRNA:miRNA* duplex representation for the hsa-mir-17 hairpin is shown in Figure 2.1B.

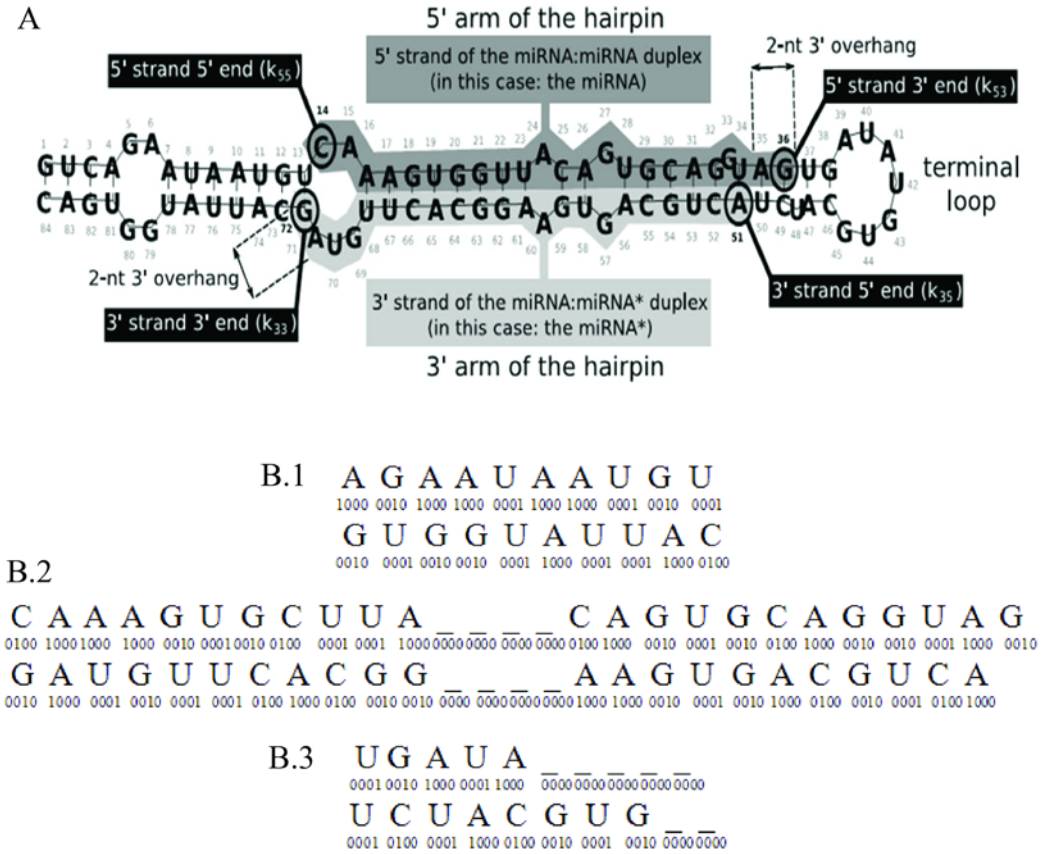

Figure 2.1. A. Anatomy of the hsa-mir-17 hairpin, showing the duplex (grey), its four ends (k55, k53, k33, k35) and the two overhangs. B. Vector representation of the true miRNA:miRNA duplex stemming from the has-mir-17. B.1. 5'strand 5'end (top) and 3'strand 3' end (bottom) 10-nts flanking region nucleotide sequence, B.2. 5' strand (top)-3' strand (bottom), duplex nucleotide sequence. Zero padding takes place in the middle. B.3. 5'strand 3'end (top), 3'strand 5' end (bottom) 10-nts flanking region nucleotide sequence. Zero padding for flanking regions takes place at the end, because the 10-nts flanking region extends beyond the 5'arm 3'end/ 3'arm 5' end , which is defined to be halfway from the terminal loop.*

Thermodynamics

As in the case of the representation of nucleotide sequence information, we used a fixed length numerical vector to encode thermodynamics information. Thermodynamics information was obtained by using the "UNAFold.pl" program [89]. "UNAFold.pl" returns a ".det" file from which we extract this information. The program assigns to each nucleotide a value which ranges from 0 to 1. If no number is assigned to a specific nucleotide we use the number which has been assigned to the closest nucleotide upstream. We construct a vector with these values which we use as input to the SVM. Note that, the process of zero padding is also used as described above, with the only difference that one zero was used for each missing nucleotide instead of four as in the case of the sequence.

Structure

Structural information was obtained by using the RNAfold program [87] with the default parameters (-p -d0 -noLP -noPS). The program returns a string which consist of dots "." and brackets "(" ")" for each hairpin. We encode this information to a vector by converting the dot – bracket notation to 0 and 1 respectively. The process of zero padding is used as described in the previous section.

Entropy

Entropy information was acquired by using RNAfold' utility: "mountain.pl", http://web.mit.edu/seven/src/ViennaRNA-1.5/Utils/ [87]. As in the case of thermodynamics the program assigns to each nucleotide a value which ranges from 0 to 1. We construct a vector with these values which we use as input to the SVM. The process of zero padding is again used as described in the thermodynamics section.

Sequences

MiRBase hairpin sequences were used in all experiments described below [90]. However, miRBase "stem-loop" entries, do not always correspond to the exact pre-miRNA sequence, but consist of the latter extended with some flanking nucleotides. Furthermore, the Drosha cleavage site is determined by its distance (~11 base pairs, 13 nts on the 5' arm and 11nts on the 3' arm) from the stem – single stranded tails junction [3]. For these reasons, the described experiments were also performed after adding 13 nts upstream and downstream the miRBase stem-loop entries. We name "sequence" the miRBase sequences without adding any extra nucleotides, and "sequence +13" the miRBase sequences plus 13nts upstream and downstream the stem-loop. The sequence of these nucleotides is not random but corresponds to the genomic sequence adjacent to the stem/loop structures.

The training and testing procedures are depicted in Figure 2.2. Briefly, given a set of training hairpins, the process consists of the following steps: (a) all entries with unknown duplexes and/or multi-branch structures (structures with more than one stems are considered multi-branch) are filtered out. (b) For each hairpin, all possible duplexes (~10,000 per hairpin) are generated and divided into the single Positive (experimentally verified duplex) and Negative (the rest) examples. To reduce training time, only 100 randomly selected negative duplexes per positive sample are used for training. (c) Selected positive and negative duplexes are used to train an SVM classifier with a full polynomial Kernel $K(x_i, x_j) = (x_i \bullet x_j + 1)^d$, where • represents the inner product of the vectors and d is the degree of the polynomial. Since the distribution of the two classes is quite unbalanced (1:100), the penalty in the SVM objective function is weighted differently for each class, namely as number-of-samples/(c*number-of-positive-samples) for positive examples and number-of-samples/(c*number-of-negative-samples), for negative examples, where c is a hyper-parameter. Thus, the Hinge loss for examples of the rare class (positives) is higher than the loss for examples of the abundant class. The SVM software used is the MATLAB interface for LIBSVM (version 3.11) [91].

Testing follows a similar procedure: multi-branch or unfoldable hairpins are first filtered out. Then, per-hairpin, all candidate duplexes are generated (note that the ranges of the overhang's and mature's length are always deduced from the

training set alone). The duplex with the maximum SVM score is selected as the algorithm's final prediction.

Prediction error is assessed using two metrics (see Text S1): (a) The ACSAE (All Corners Sum Absolute Error) is the sum of absolute errors in number of nucleotides from true position between the actual and the predicted duplex end, taken over all four ends of the duplex. (b) The EAE (End Absolute Error) focuses on a specific end of the duplex; it is the absolute error of the predicted minus the true position (in nucleotides) in a specific duplex end. For example, if the true positions are *k55* = XX_{55}, *k53* = XX_{53}, *k35*= XX_{35} and *k33* = XX_{33}, and the predicted positions are YY_{55}, YY_{53}, YY_{35}, YY_{33} respectively, then the ACSAE on this duplex is ACSAE = ($|XX_{55} - YY_{55}| + |XX_{53} - YY_{53}| + |XX_{35} - YY_{35}| + |XX_{33} - YY_{33}|$). The EAE for each duplex end is $|XX_{55} - YY_{55}|$, $|XX_{53} - YY_{53}|$, $|XX_{35} - YY_{35}|$, $|XX_{33} - YY_{33}|$, respectively.

To measure prediction accuracy, we define as "correct" a prediction with error less or equal to a number *x*. Then, the prediction accuracy for an error bound of at most *x*, denoted as Accu(*x*), is the percentage of correct predictions in the test set. For example, if a model identifies correctly the position of 50% of duplexes with ACSAE ≤ 4, it has accuracy at 4nt of 50%: Accu(4) = 0.5.

Statistical significance of the results is assessed by assuming the null hypothesis that two methods have the same accuracy for a given error bound and applying the Fisher's exact test.

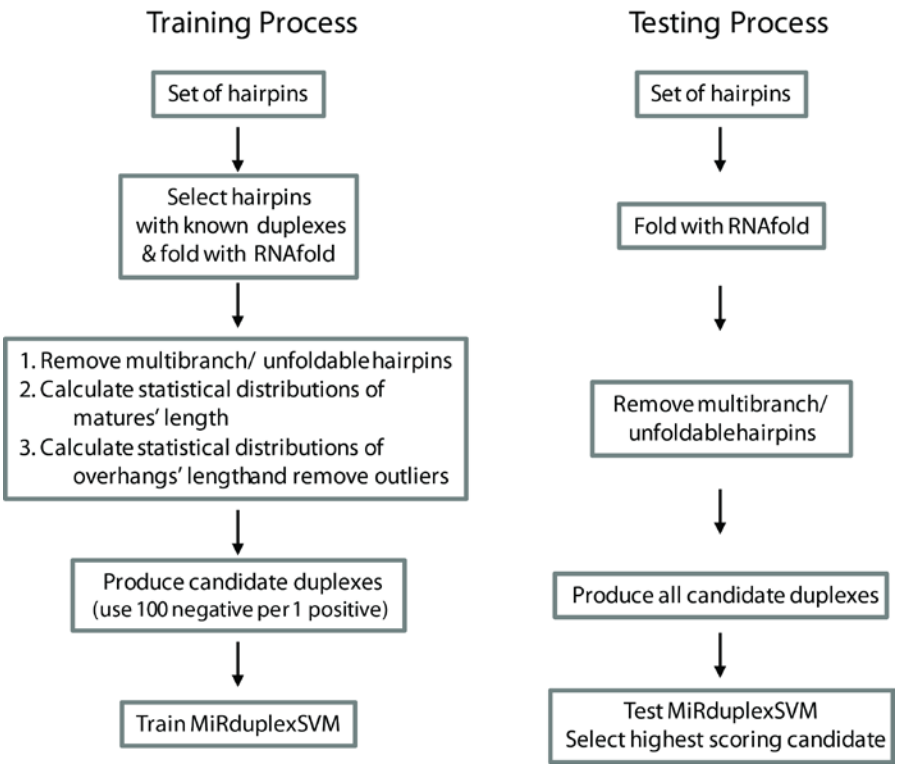

Figure 2.2. Flowcharts of the training and testing procedures.

Optimization of SVM Hyperparameters

The method has three hyper-parameters to optimize: the cost *c*, the degree of the kernel *d* of the SVM, and the length of the flanking region *l* (number of nucleotides before and after the duplex) of the vector representation. We note that proper selection of these hyper-parameters was critical to achieving high performance. They were optimized once using 5-fold cross validation on a

randomly-selected subset of version 17.0 of miRBase, consisting of 70% (658 in number) human/mouse hairpins with known duplexes. The values tested were: d = 1, 2, 3, c = 100, 10, 1, 0.1, 0.01, 0.001 and l = 0, 3, 6, 9, 10…15 nts. The performance during cross-validation was measured in terms of predicting the exact location of the duplex by calculating the sum of the absolute error taken over all four ends of the duplex (ACSAE). The best performing combination of parameters found was d = 3, c = 0.01 and l = 10nts. This combination of parameters was employed in all MiRduplexSVM models reported here. *To ensure unbiased estimations of performance, the 658 hairpins used for hyper-parameter optimization were excluded from all test sets used in subsequent evaluations.*

Producing a Simple Geometric Locator as a Baseline Comparison

We also develop a Simple Geometric Locator (S.G.L.), whereby each of the four ends of the duplex is predicted by its average location in a training set. Specifically, for any given hairpin, the location of each of the four ends of known duplexes is found by calculating its distance from the tip of the terminal loop. This is done for all hairpins in the training set and the average distances (rounded to the closest integer) are then used to generate the predictions of the S.G.L. for any new hairpin in the test set. The terminal loop tip was chosen as the reference point as it does not depend on the length of the pre-miRNA flanking regions included in the hairpin sequence.

RESULTS

Model selection

Extensive experimentation was first performed in order to find the minimum set of features like sequence, structure, entropy, thermodynamics or distance based characteristics needed to obtain maximum accuracy (see Table 2.1). Based on the error metric ACSAE we find two models that fulfill these requirements, model 8 and model 11, as shown in Figure 2.3. Model 8 has been trained and tested using sequence and entropy information while model 11 using only sequence information (see Table2.1). Both models achieve similar accuracy in finding the four corners of the miRNA duplex, for an error tolerance of up to 3nts. In addition they perform better or similar to other models which are more complex. While model 8 achieves better ACSAE for errors higher than 3 nts, model 11 is simpler and has a higher probability of producing the true duplex. We thus optimized the parameters of both models. Specifically, for both algorithms the degree of the kernel and the cost parameter of the SVM as well as the flanking sequence length (number of nucleotides before and after the duplex) were optimized using five-fold cross validation. In both cases the best performing parameters were found to be: d = 3, c = 0.01 and l = 10. The mean prediction accuracies of the optimized models versus the ACSAE are shown in Figure 2.4. Even though, the "Sequence – Entropy" model seems to have better performance than the "Sequence" model, the observed differences were not statistically significant and we decided to select the simplest model.

TABLE 2.1 SET OF FEATURES USED IN MODEL SELECTION

	miRBase sequences		**Candidate Production**		**SVM input**												
Model Number	Sequence	Sequence + 13	All	Selected	Sequence	Entropy	Thermodynamics	Structure	Flank	A	B	C	D	E	F	G	H
1		+		+	+	+											
2		+		+	+												
3		+	+		+												
4		+	+		+	+											
5		+	+		+		+										
6	+		+		+			+	12	+	+	+	+				
7	+		+		+			+	12	+	+	+	+	+	+	+	+
8	+			+	+	+			13								
9	+			+	+		+		13								
10	+			+	+				13								
11	+		+		+				13								
12	+		+				+		12	+	+	+	+	+	+	+	+
13	+			+			+		12								

Features used for model selection. In all models sequences from miRBase were used. "Sequence" means the exact miRBase sequences were used. "Sequence + 13" means miRBase sequences were used after adding 13nts upstream and downstream each miRBase stem-loop entry. Candidate duplexes were also produced in two ways, "All" and "Selected" (see text for details). The SVM input consisted of different features, which were used alone or in combination with others. Sequence, entropy, thermodynamics, structure and several distance-based features were used. Flank: the flanking sequence at both ends of each duplex strand in the representation of a candidate duplex. A: miRNA length, 5p strand, B: miRNA length, 3p strand, C: overhang length, 5p strand, D: overhang length, 3p strand, E: distance k55 from the loop tip, F: distance k53 from the loop tip, G: distance k35 from the loop tip, H: distance k33 from the loop tip, I: hairpin loop sequence length, J: distance k53 from the start of the loop, K:distance k35 from the end of the loop. + means that this feature was included in SVM training and testing.

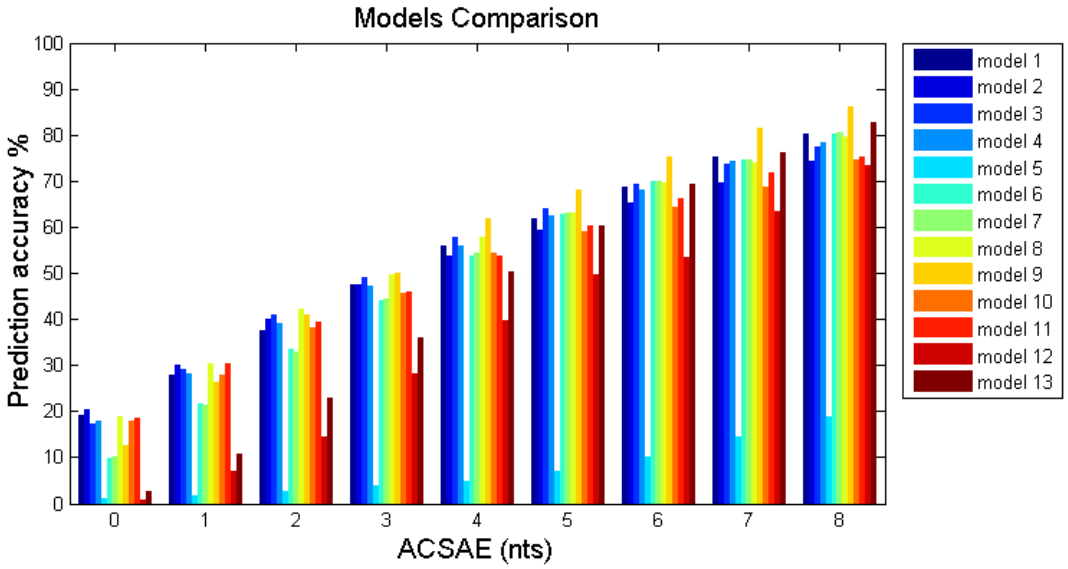

Figure 2.3. Mean prediction accuracies achieved by the different models. The mean prediction accuracy shown in the figure was calculated in the following way. During each five-fold cross validation, 5 models were produced M1, M2, M3, M4, M5. For each model we calculate its prediction accuracy PA1, PA2, PA3, PA4 and PA5 and average all of them to obtain the mean prediction accuracy, MPA, MPA= (PA1 + PA2 + PA3 + PA4 + PA5)/ 5. The simplest models showing good performance for low error values are model 8 and model 11.

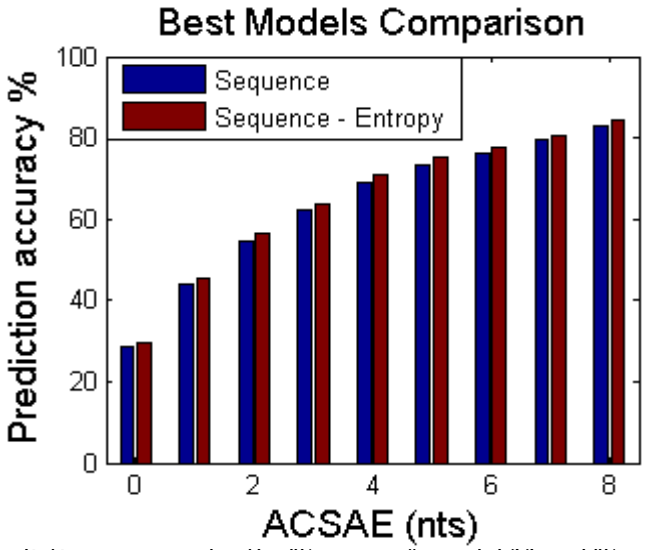

Figure 2.4. Mean prediction accuracy for the "Sequence"- model 11 and "Sequence - Entropy" – model 8. The prediction accuracy shown in the figure was calculated as in Figure 2.3. Prediction accuracies were obtained using the best performing combination of the degree of the kernel d, the cost c and the flanking sequence length l. The "Sequence – Entropy" model has better performance than the "Sequence" model, however by performing Wilcoxon "ranksum" tests we note that the observed *difference was not statistically significant.*

Comparison with a Simple Geometric Locator

The performance of MiRduplexSVM was first compared to that of the Simple Geometric Locator. Both methods were trained on the dataset used to optimize the hyper-parameters of MiRduplexSVM and tested on the remaining 30% of hairpins with known duplexes (290 hairpins) in version 17.0 of miRBase. Figure 2.6A shows the prediction accuracy of each tool as a function of the ACSAE while Figure 2.6G (blue line) and Figure 2.7 (blue lines) show the prediction accuracy of MiRduplexSVM against that of the S.G.L. estimated using the ACSAE (0-8nts, Fig. 3G) or the EAE (0–5nts, Fig. 4), respectively. In all cases MiRduplexSVM greatly outperforms the S.G.L., especially for small error values. The observed difference in performance is statistically significant for ACSAE of 0-15 nucleotides (Table 2.2 line 6, p=0.05), and for EAE of 0-3 nucleotides (Table 2.3, lines 21 – 24, p=0.05), beyond which both methods behave similarly. Having shown that the distance from the tip loop, while a very simple approach, is not sufficient to identify miRNA duplexes, we next compare MiRduplexSVM with existing miRNA mature prediction tools.

Comparison with other State of the Art Tools

In the following paragraphs we compare MiRduplexSVM with four state-of-the-art mature miRNA prediction tools, namely MatureBayes [55], MiRPara [56],

MaturePred [57] and MiRdup [58]. To ensure fairness, in each comparison MiRduplexSVM is trained with the original training set of the compared tool and evaluated on a common hold-out test set. The procedure for building the various test sets is depicted in Figure 2.5. Briefly, we start with all hairpins in miRBase version 19.0 and exclude the ones previously seen by any of the compared tools (during training or parameter optimization), resulting in ~16.000 hairpins. This set, hereby termed "Test Set A" is used to extract test sets for the various comparisons according to the specifications of each tool. Test Set A.1 was used to evaluate MatureBayes and MiRPara and was generated by randomly selecting ~5.000 hairpins, while maintaining the same species ratio of the original test set A, out of which ~2.500 had known duplexes. Test Set A.2 consisted of hairpins (2688, 1578 with known duplexes) from Test Set A.1 which belonged to the species used in the original test set of MaturePred_Mammals. Comparison with

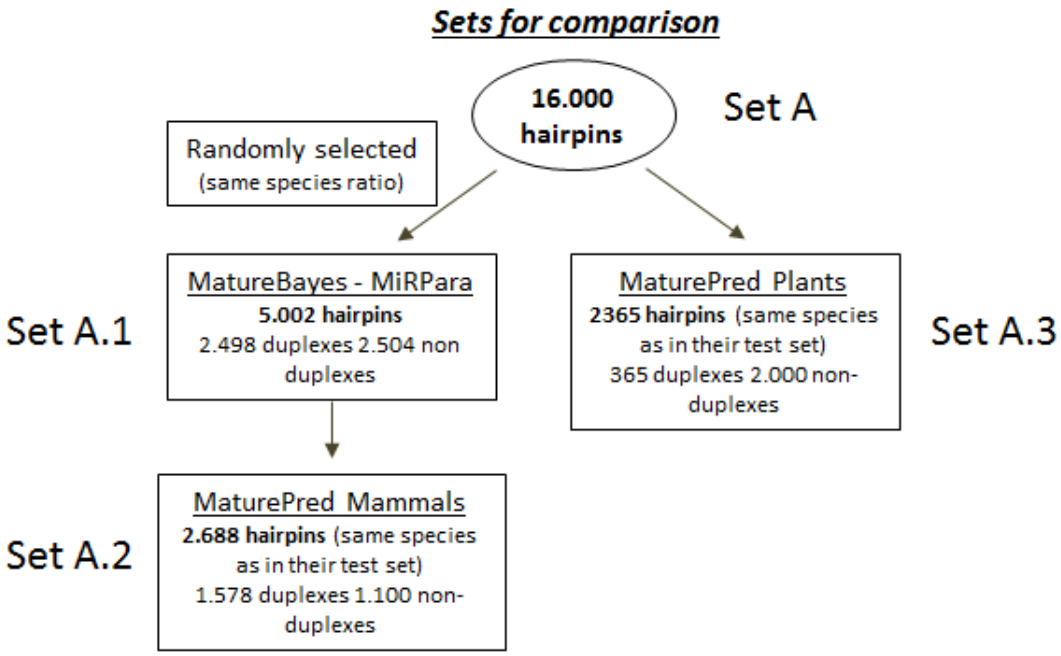

Figure 2.5. Building test sets procedure.

MaturePred was also done for plants, using the Test Set A.3. This set consisted of all plant hairpins from Test Set A that belonged to the species used in the original test set of MaturePred_Plants (2365, 365 with known duplexes).

Performance accuracies on (a) duplex identification, using the ACSAE, and (b) independent corner identification, using the EAE were estimated only on hairpins that are predicted to contain a mature miRNA by both of the compared tools. Comparisons shown in Figures 2.6 and 2.7 are performed by finding the prediction accuracy of each tool for an ACSAE of 0-8nts and an EAE of 0-5nts and (a) plotting these accuracies as a function of the ACSAE metric (Figure 2.6A-F) or (b) plotting these accuracies against each other (Figure 2.6G and Figure 2.7). Specifically, on Figure 2.6G the prediction accuracy, measured as the ACSAE (in %), of MiRduplexSVM (y axis) at error points 0-8nts is plotted against the respective accuracy of each compared tool (x-axis). For example, the first point (triangle) on each line represents the pair of accuracies ($Accu_i(0)$, $Accu_{MiRduplexSVM}(0)$), the second point (rhombus), the pair ($Accu_i(1)$, $Accu_{MiRduplexSVM}(1)$) and so on, where *i* is the compared tool. The points that correspond to comparisons against a given tool *i* are connected with a line. Thus, if a line is right on the diagonal, then the two methods achieve the same accuracy for the same error tolerance. If it is above the diagonal, then MiRduplexSVM achieves the same accuracy for smaller error levels than the method compared against. The same applies to Figure 2.7, with the only difference that the accuracy is measured using the EAE instead of the ACSAE. The results of each tool comparison are summarized below.

MatureBayes [55] was the first tool specifically developed to address the problem of mature miRNA identification. It utilizes a Naive Bayes classifier to identify mature miRNA candidates based on sequence and secondary structure information of their miRNA precursors [55]. It generates one prediction per strand and uses the 2nts overhang rule to define the predicted miRNA* on the opposite strand, creating two hypothetical duplexes. For this comparison, MiRduplexSVM was trained with the original training set of MatureBayes and both algorithms were evaluated on Test-Set A1. MatureBayes' predictions were obtained using the downloadable version of the algorithm. The ACSAE was applied on MatureBayes' hypothetical duplex which obtained the highest score. The EAE were in turn applied on the strand-specific molecules of the highest scoring candidate duplex for both tools. Comparison to MiRduplexSVM is shown in Figures 3B and 3G (pink line) for duplex prediction and Figure 2.7 (pink lines) for independent corner prediction. As evident from the figures, MiRduplexSVM significantly outperforms MatureBayes (pink lines are above the diagonal) in both duplex (up to 12 nts, Table 2.2, row 1) and independent corner (up to 4nts, Table 2.3, rows 1–4) prediction.

MiRPara [56] is an SVM-based tool for mature miRNA prediction. We compare MiRduplexSVM with the stand alone application, MiRPara 4.2, which was available at the time of evaluation. We trained our algorithm with all duplexes in miRBase version 13.0, which was used as a training set for MiRPara 4.2, and both algorithms were tested on Test Set A.1 (Figure 2.7). MiRPara generates several independent predictions for each strand of any given hairpin, whereas MiRduplexSVM predicts only one duplex per hairpin. In order to compare the

two tools in terms of the ACSAE we needed to produce a single hypothetical duplex for MiRPara. To this end, we consider the highest scoring predictions per strand as the two sides of the hypothetical duplex of any given hairpin. For comparisons using the EAE, we contrasted the top scoring prediction of MiRPara per strand against the prediction of MiRduplexSVM for the same strand. Strand-specific comparisons were performed independently of one another, i.e. there was no requirement that both strands produce a mature miRNA. Comparison to MiRduplexSVM is shown in Figures 3C and 3G (cyan line) for duplex prediction and Figure 2.7 (cyan lines) for independent corner prediction. It should be noted that MiRPara gave a prediction for only 3774 out of the 5000 hairpins used for testing. Prediction accuracy for both models was calculated on these 3774 hairpins, which biases the comparison in favour of MiRPara. As evident from the figures, MiRduplexSVM significantly outperforms MiRPara (cyan lines are above the diagonal) in both duplex (up to 8nts, Table 2.2, row 2) and independent corner (up to 2nts, Table 2.3, rows 5-8) prediction. Note, that the S.G.L. also has a good performance for errors beyond 3-4nts (Table 2.3) indicating that even the simplest method can find the true mature when the tolerance for errors is more than a couple of nucleotides per corner.

MaturePred [57] employs an SVM classifier to predict the region which is most likely to contain the mature miRNA molecule in each strand of a hairpin. It consists of two models, one specialized in plants, hereby termed MaturePred_Plants and one specialized in mammals, hereby named MaturePred_Mammals. We compare MiRduplexSVM with each model separately. In each case we train MiRduplexSVM with the respective MaturePred's training

set and evaluate performances on Test Sets A.2 (for mammals) and A.3 (for plants). In both comparisons, MaturePred's predictions were acquired by using the online version of the respective model following the recommendations on its web site. As in the case of MiRPara, MaturePred gives multiple independent predictions per strand. Thus, we measure the ACSAE and EAE for each corner as described for MiRPara. Comparison to MiRduplexSVM is shown in Figures 3D (Plants), 3E (Mammals) and 3G (Plants: green line, Mammals: red line) for duplex prediction and Figure 2.7 (Plants: green lines, Mammals: red lines) for independent corner prediction. As evident from figures 3D,E,G (lines above diagonal), MiRduplexSVM significantly outperforms MaturePred on duplex prediction for both plant and mammalian hairpins (for all errors tested, Table 2.2, rows 3 & 4). For independent corner prediction, MiRduplexSVM outperforms MaturePred in a statistically significant manner only for mammalian hairpins (red lines above diagonal), while for plant hairpins both tools achieve similar performances (Fig. 4, green lines on the diagonal and Table 2.3, rows 9 - 16).

MiRdup [58] is the latest tool that tackles the problem of mature miRNA identification. It does so by finding the most likely miRNA location within a given pre-miRNA. MiRdup is based on a random forest classifier trained with experimentally validated miRNAs from miRbase, with features that characterize the miRNA–miRNA* duplex. MiRdup predicts the most probable miRNA duplex on a given hairpin. Both MiRduplexSVM and MiRdup were trained on 70% of miRBase 17.0 (658 hairpins) and tested on the remaining 30% (290 hairpins). This was done since the MiRdup's downloadable model was trained on the entire miRBase 19.0, and a lot of errors occurred when we tried to use it. Comparison to

MiRduplexSVM is shown in Figures 3F and 3G (purple line) for duplex prediction and Figure 2.7 (purple lines) for independent corner prediction. As evident from the figures, MiRduplexSVM significantly outperforms MiRdup (purple lines are above the diagonal) in both duplex (up to 15nts, Table 2.2, row 5) and independent corner (up to 3nts, Table 2.3, rows 17 – 20) prediction.

In sum, on the task of duplex prediction, MiRduplexSVM outperforms all other tools it has been compared to for an error tolerance of at least 8nts and the increase in performance accuracy ranges from ~10% to 60% (Figure 2.6G and Table 2.2). With respect to individual end comparisons (Figure 2.7 and Table 2.3), MiRduplexSVM is again found to outperform all methods, particularly for small EAEs (0-4nts). The only exception is MaturePred-Plants which achieves a similar performance. The latter maybe due to the parameter optimization of MiRduplexSVM which was done using mammalian hairpins and/or the small number of plant hairpins used to train MiRduplexSVM (198) compared to MaturePred_Plants (1.323).

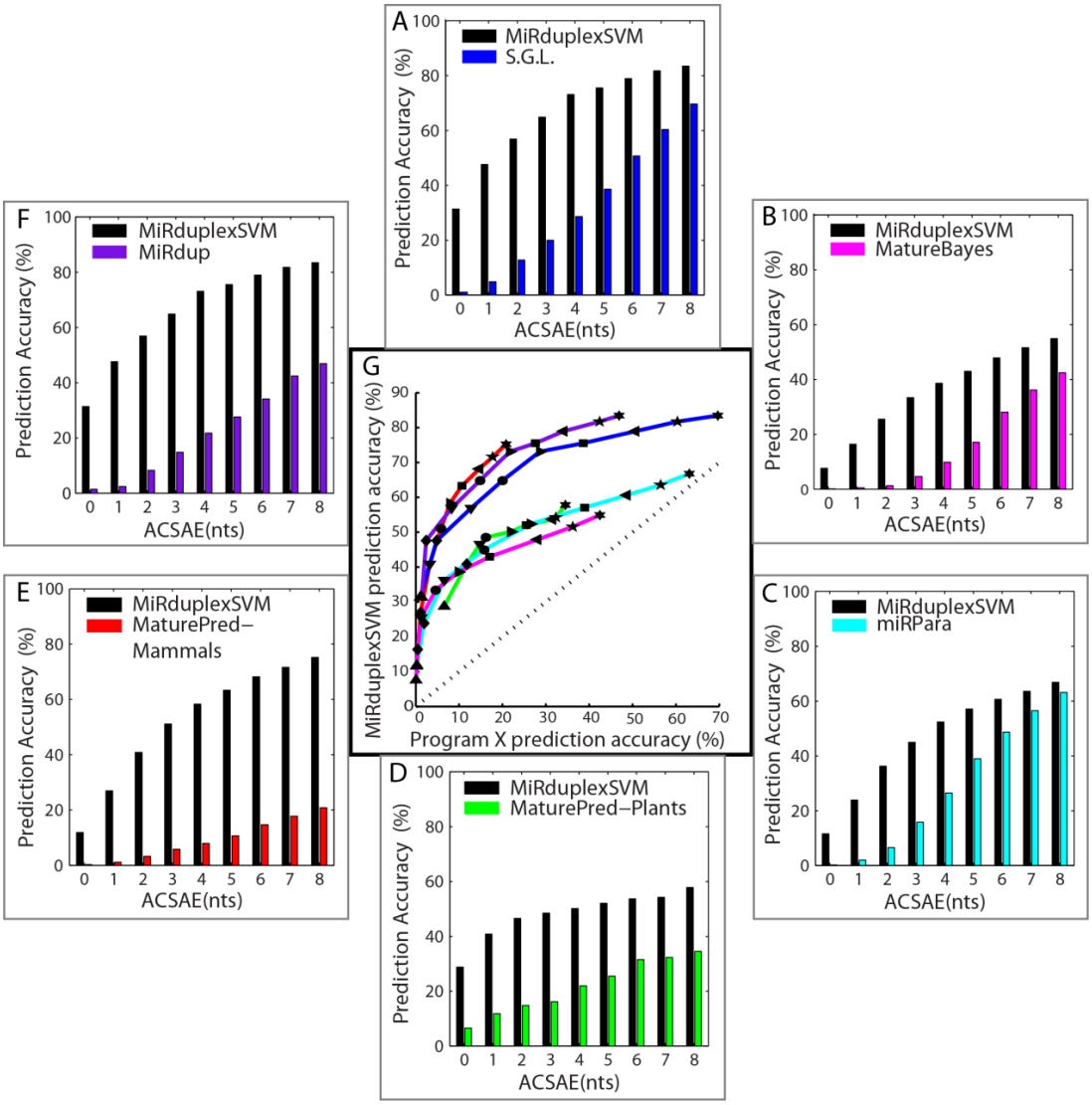

Figure 2.6. Prediction accuracy of MiRDuplexSVM and six other methods on duplex identification. Panels A-F show the prediction accuracy (y-axis) of MiRDuplexSVM (in black) and a second compared tool (in colour) as a function of the All Corners Sum Absolute Error (ACSAE, x axis) for errors of 0-8nts. The performance of the Simple Geometric Locator (S.G.L.), MatureBayes, miRPara, MaturePred-Plants, MaturePred-Mammals and MiRdup is shown in A – blue bars, B – pink bars, C – cyan bars, D – green bars, E -red bars and F -purple bars, respectively. Panel G shows the prediction accuracy of MiRDuplexSVM (y axis) against the prediction accuracy of each compared tool (x axis). The colour code is the same as in A-F. Symbols (upward triangle, diamond, downward triangle, circle, right pointed triangle, square, left pointed triangle, pentagram star and hexagram star) correspond to errors less than or equal to 0, 1, 2, 3, 4, 5, 6, 7, 8 nucleotides, respectively. All points above the diagonal in G are statistically significant at level 0.05.

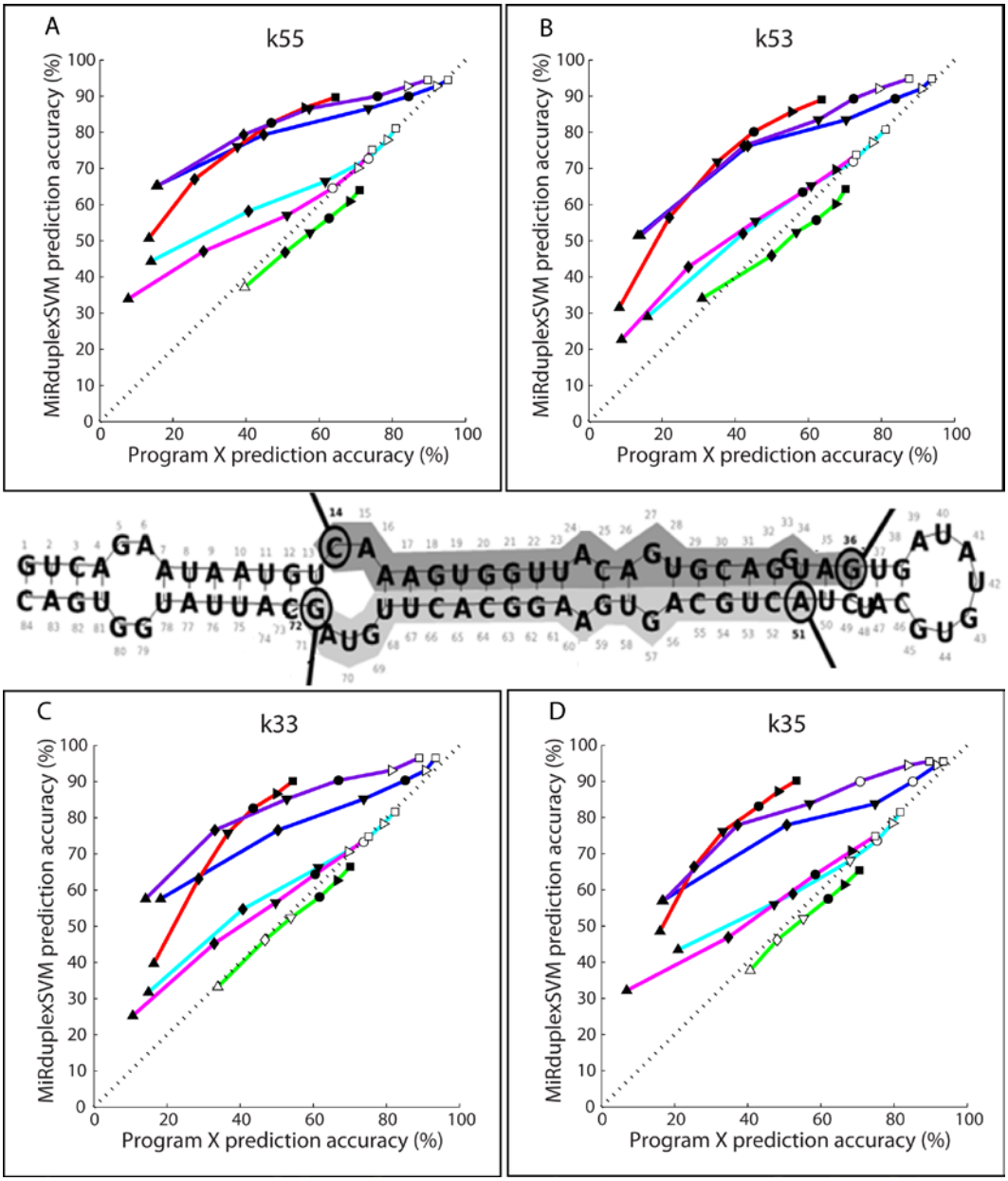

Figure 2.7. Prediction accuracy of MiRDuplexSVM versus six other methods on corner identification. Performance accuracies are estimated using the EAE for errors of 0-5nts. In each panel, the y axis shows the prediction accuracy of MiRduplexSVM (in %) and the x axis shows the prediction accuracy of other methods (in %). The colour and symbol scheme is the same as in Figure 2.6G. Statistically significant results are indicated with filled symbols.

TABLE 2.2 PREDICTION ACCURACIES, UP TO 20 NTS DEVIATION

All Corners Sum Absolute Error in nucleotides																				
	<= 0	<= 1	<= 2	<= 3	<= 4	<= 5	<= 6	<= 7	<= 8	<= 9	<= 10	<= 11	<= 12	<= 13	<= 14	<= 15	<= 16	<= 17	<= 18	<= 19	<= 20
MiRduplexSVM/ MatureBayes - Prediction accuacy %	7.61 %/ 0.08 % ***	16.38 %/ 0.44 % ***	25.52 %/ 1.25 % ***	33.41 %/ 4.59 % ***	38.61 %/ 9.86 % ***	42.95 %/ 17.07 % ***	47.91 %/ 28.1 % ***	51.57 %/ 36.19 % ***	54.91 %/ 42.47 % ***	57.25 %/ 47.91 % ***	59.86 %/ 54.15 % ***	62.04 %/ 57.29 % ***	63.81 %/ 61.27 % *	65.66 %/ 64.29 % ns	67.11 %/ 67.15 % ns	68.72 %/ 68.84 % ns	70.21 %/ 71.3 % ns	71.46 %/ 72.46 % ns	72.95 %/ 73.95 % ns	73.95 %/ 74.64 % ns	74.96 %/ 75.6 % ns
MiRduplexSVM/ MiRPara - Prediction accuacy %	11.58 %/ 0.21 % ***	23.89 %/ 1.95 % ***	36.26 %/ 6.58 % ***	44.95 %/ 15.79 % ***	52.37 %/ 26.47 % ***	57.05 %/ 38.95 % ***	60.63 %/ 48.68 % ***	63.58 %/ 56.47 % ***	66.79 %/ 63.11 % **	69.26 %/ 68.11 % ns	70.74 %/ 72.11 % ns	72.79 %/ 74.58 % ns	74.63 %/ 76.32 % ns	76.11 %/ 78% ns	77.26 %/ 79.16 % ns	78.68 %/ 80.16 % ns	80.21 %/ 81.16 % ns	80.84 %/ 82.05 % ns	81.89 %/ 83% ns	82.79 %/ 83.74 % ns	83.53 %/ 84% ns
MiRduplexSVM/ MaturePred_Plants - Prediction accuacy %	28.77 %/ 6.58 % ***	40.82 %/ 11.78 % ***	46.58 %/ 14.79 % ***	48.49 %/ 16.16 % ***	50.14 %/ 21.92 % ***	52.05 %/ 25.48 % ***	53.7 %/ 31.51 % ***	54.25 %/ 32.33 % ***	57.81 %/ 34.52 % ***	58.63 %/ 36.71 % ***	59.45 %/ 39.73 % ***	59.73 %/ 41.37 % ***	61.64 %/ 43.56 % ***	63.56 %/ 45.75 % ***	64.93 %/ 50.41 % ***	65.75 %/ 51.51 % ***	67.67 %/ 52.88 % ***	67.67 %/ 53.42 % ***	68.49 %/ 55.62 % ***	69.32 %/ 56.99 % ***	69.86 %/ 58.9 % **
MiRduplexSVM/ MaturePred_Mammals - Prediction accuacy %	11.85 %/ 0.25 % ***	26.93 %/ 1.14 % ***	40.87 %/ 3.23 % ***	51.08 %/ 5.83 % ***	58.3 %/ 7.92 % ***	63.31 %/ 10.65 % ***	68.19 %/ 14.7 % ***	71.61 %/ 17.74 % ***	75.22 %/ 20.85 % ***	77.19 %/ 23.07 % ***	79.66 %/ 25.48 % ***	81.18 %/ 27.44 % ***	82.83 %/ 29.85 % ***	84.41 %/ 31.94 % ***	85.61 %/ 33.84 % ***	86.76 %/ 35.8 % ***	87.45 %/ 37.83 % ***	88.21 %/ 40.18 % ***	88.97 %/ 42.65 % ***	89.92 %/ 44.8 % ***	90.49 %/ 47.34 % ***
MiRduplexSVM/ MiRdup - prediction accuracy %	31.38 %/ 1.03 % ***	47.59 %/ 4.83 % ***	56.9 %/ 12.76 % ***	64.83 %/ 20% ***	73.1 %/ 28.62 % ***	75.52 %/ 38.62 % ***	78.97 %/ 50.69 % ***	81.72 %/ 60.34 % ***	83.45 %/ 69.66 % ***	86.55 %/ 74.48 % ***	88.62 %/ 78.97 % **	88.97 %/ 80.34 % **	90.69 %/ 81.38 % ***	91.38 %/ 83.1 % **	91.72 %/ 85.86 % *	92.76 %/ 88.28 % *	93.79 %/ 91.72 % ns	94.14 %/ 92.07 % ns	94.14 %/ 93.1 % ns	94.48 %/ 93.45 % ns	95.86 %/ 93.45 % ns
MiRduplexSVM/ Simple Geometric Locator - Prediction accuacy %	31.38 %/ 1.03 % ***	47.59 %/ 4.83 % ***	56.9 %/ 12.76 % ***	64.83 %/ 20% ***	73.1 %/ 28.62 % ***	75.52 %/ 38.62 % ***	78.97 %/ 50.69 % ***	81.72 %/ 60.34 % ***	83.45 %/ 69.66 % ***	86.55 %/ 74.48 % ***	88.62 %/ 78.97 % **	88.97 %/ 80.34 % **	90.69 %/ 81.38 % ***	91.38 %/ 83.1 % **	91.72 %/ 85.86 % *	92.76 %/ 88.28 % *	93.79 %/ 91.72 % ns	94.14 %/ 92.07 % ns	94.14 %/ 93.1 % ns	94.48 %/ 93.45 % ns	95.86 %/ 93.45 % ns

*The sum of the absolute error taken over all four ends of the predicted (MiRduplexSVM) or the hypothetical (MaturePredPlants, MaturePredMammals, MiRPara, MatureBayes) duplexes is calculated. MiRduplexSVM has been trained on each program's training set and their performance has been accessed on a common blind test set. Fisher exact test were performed to examine if the observed differences are statistical significant. *** corresponds to pvalue ≤0.001, ** to pvalue ≤0.01, * to pvalue ≤0.05, and ns to non statistical.*

TABLE 2.3 PREDICTION ACCURACIES, UP TO 8 NTS DEVIATION.

		End Absolute Error (EAE)								
		<= 0	<= 1	<= 2	<= 3	<= 4	<= 5	<= 6	<= 7	<= 8
MiRduplexSVM/ MatureBayes - Prediction accuracy %	k55	33.93% / 7.79% ***	47.07% / 28.34% ***	57.03% / 51.19% ***	64.58% / 63.68% ns	70.14% / 70.58% ns	75.19% / 74.48% ns	78.53% / 77.44% ns	80.84% / 79.61% ns	83.44% / 81.35% *
	k53	22.72% / 8.96% ***	42.75% / 27.23% ***	55.46% / 45.55% ***	63.46% / 58.44% ***	69.65% / 67.54% *	73.78% / 73.13% ns	77.55% / 76.74% ns	80.67% / 79.72% ns	82.84% / 81.27% *
	k35	32.15% / 6.9% ***	46.85% / 34.66% ***	56%/ 47.2% ***	64.23% / 58.49% ***	70.73% / 68.4% *	74.88% / 74.85% ns	77.58% / 78.85% ns	80.49% / 80.75% ns	82.63% / 82.63% ns
	k33	25.28% / 10.66% ***	45.21% / 32.89% ***	56.5%/ 49.68% ***	64.36% / 60.47% ***	70.6%/ 69.62% ns	74.8%/ 75.12% ns	77.71% / 79.16% ns	80.22% / 81.31% ns	83.1%/ 83.18% ns
MiRduplexSVM/ MiRPara - Prediction accuracy %	k55	44.34% / 14.03% ***	58.2%/ 40.65% ***	66.45% / 61.68% ***	72.75% / 73.44% ns	77.86% / 78.59% ns	81.07% / 80.93% ns	83.95% / 82.6% ns	85.62% / 84.79% ns	87.54% / 85.66% *
	k53	28.99% / 16.05% ***	51.97% / 42.15% ***	65.23% / 60.81% ***	71.91% / 72.26% ns	77.2%/ 77.62% ns	80.75% / 81.17% ns	83.85% / 82.81% ns	85.49% / 83.99% ns	87.16% / 85.03% *
	k35	43.43% / 20.95% ***	58.9%/ 52.3% ***	68.08% / 67.98% ns	73.6%/ 75.32% ns	78.44% / 79.55% ns	81.52% / 81.72% ns	83.65% / 83.38% ns	85.14% / 84.77% ns	87.14% / 86.15% ns
	k33	31.75% / 14.93% ***	54.71% / 40.72% ***	66.22% / 61.37% ***	73.32% / 73.73% ns	78.3%/ 79.18% ns	81.65% / 82.36% ns	84.12% / 84.5% ns	85.65% / 85.78% ns	87.31% / 86.97% ns
MiRduplexSVM/ MaturePred_Plants - Prediction accuracy %	k55	37.14% / 39.62% ns	46.83% / 50.69% *	52.22% / 57.47% **	56.23% / 62.71% ***	60.89% / 68.32% ***	64.02% / 71.01% ***	66.86% / 74.58% ***	69.85% / 76.84% ***	72.1%/ 78.88% ***
	k53	34.09% / 30.95% *	45.88% / 49.96% *	52.37% / 56.74% *	55.72% / 62.2% ***	60.16% / 67.52% ***	64.31% / 70.21% ***	66.28% / 73.71% ***	70.07% / 77.2% ***	72.54% / 78.51% ***
	k35	37.73% / 40.68% ns	46.13% / 48.05% ns	52.03% / 55.2% ns	57.48% / 61.97% **	61.39% / 66.4% **	65.44% / 70.52% **	69.57% / 73.69% **	72.37% / 76.49% **	74.72% / 78.78% **
	k33	33.24% / 33.9% ns	46.2%/ 46.79% ns	52.32% / 53.87% ns	58.07% / 61.75% *	62.64% / 66.54% *	66.47% / 70.08% *	69.2%/ 73.25% *	72.37% / 76.27% *	75.31% / 79% *
MiRduplexSVM/ MaturePred_Mammals - Prediction	k55	50.69% / 13.44% ***	67.03% / 25.94% ***	75.96% / 37.62% ***	82.61% / 46.94% ***	86.79% / 56.1% ***	89.69% / 64.47% ***	92.11% / 69.69% ***	93.35% / 74.11% ***	94.77% / 80.33% ***

accuracy %	k53	31.54% / 8.31% ***	56.44% / 22% ***	71.83% / 35.11% ***	80.19% / 45.13% ***	85.65% / 55.49% ***	89.03% / 63.66% ***	91.31% / 69.12% ***	93.02% / 74.16% ***	94.54% / 80.1% ***
	k35	48.56% / 16.03% ***	66.36% / 25.3% ***	76.14% / 33.23% ***	83.09% / 42.96% ***	87.21% / 48.33% ***	90.22% / 53.29% ***	92.12% / 57.78% ***	93.74% / 62.23% ***	95%/ 66.91% ***
	k33	39.67% / 16.4% ***	63.11% / 28.64% ***	75.72% / 36.61% ***	82.62% / 43.56% ***	86.61% / 49.95% ***	90.13% / 54.36% ***	92.68% / 59.64% ***	93.74% / 64.32% ***	95.04% / 69.23% ***
MiRduplexSVM/ MiRdup - Prediction accuracy %	k55	65.17% / 15.52% ***	79.31% / 44.83% ***	86.55% / 73.45% ***	90%/ 84.48% *	92.76% / 92.07% ns	94.48% / 95.17% ns	96.55% / 96.21% ns	97.93% / 97.59% ns	98.62% / 97.59% ns
	k53	51.38% / 13.45% ***	76.21% / 43.45% ***	83.45% / 70.34% ***	89.31% / 83.79% *	92.07% / 91.03% ns	94.83% / 93.79% ns	97.59% / 95.52% ns	98.28% / 97.93% ns	98.97% / 97.93% ns
	k35	56.9%/ 16.55% ***	77.93% / 50.69% ***	83.79% / 74.83% **	90%/ 85.17% ns	94.48% / 91.72% ns	95.52% / 93.45% ns	96.21% / 94.83% ns	97.59% / 95.86% ns	97.59% / 97.24% ns
	k33	57.59% / 18.28% ***	76.55% / 50.34% ***	85.17% / 73.79% ***	90.34% / 85.17% *	93.1%/ 90.69% ns	96.55% / 93.45% ns	96.55% / 94.48% ns	97.59% / 96.21% ns	97.93% / 97.59% ns
MiRduplexSVM/ Simple Geometric Locator - Prediction accuracy %	k55	65.17% / 15.52% ***	79.31% / 44.83% ***	86.55% / 73.45% ***	90%/ 84.48% *	92.76% / 92.07% ns	94.48% / 95.17% ns	96.55% / 96.21% ns	97.93% / 97.59% ns	98.62% / 97.59% ns
	k53	51.38% / 13.45% ***	76.21% / 43.45% ***	83.45% / 70.34% ***	89.31% / 83.79% *	92.07% / 91.03% ns	94.83% / 93.79% ns	97.59% / 95.52% ns	98.28% / 97.93% ns	98.97% / 97.93% ns
	k35	56.9%/ 16.55% ***	77.93% / 50.69% ***	83.79% / 74.83% **	90%/ 85.17% ns	94.48% / 91.72% ns	95.52% / 93.45% ns	96.21% / 94.83% ns	97.59% / 95.86% ns	97.59% / 97.24% ns
	k33	57.59% / 18.28% ***	76.55% / 50.34% ***	85.17% / 73.79% ***	90.34% / 85.17% *	93.1%/ 90.69% ns	96.55% / 93.45% ns	96.55% / 94.48% ns	97.59% / 96.21% ns	97.93% / 97.59% ns

*The absolute error for each one of the four ends of the duplex is calculated independently. MiRduplexSVM has been trained on each program's training set and their performance has been accessed on a common blind test set. Fisher exact test were performed to examine if the observed differences are statistical significant. *** corresponds to pvalue ≤0.001, ** to pvalue ≤0.01, * to pvalue ≤0.05 and ns to non statistical.*

Having established the superiority of our algorithm compared to existing tools, we generated a final model using all hairpins with known duplexes (5.248) available in miRBase 19.0 (latest version). The model was evaluated on 5.000 randomly selected hairpins having the same species ratio as the remaining 15.500. Mature sequences in these 5.000 hairpins were equally distributed in both strands. The accuracy of MiRduplexSVM was evaluated using the EAE since the ACSAE cannot be computed without knowledge of the true duplex. It was found to reach 55%, 39%, 54% and 43% correct prediction at 0 nucleotides deviation for k55, k53, k35 and k33, respectively (see Table 2.4). This final model achieves higher performance than the one seen in the comparisons with other tools, presumably because it is trained with a much larger training set. The model is available for download at http://139.91.171.154/duplexsvm/.

TABLE 2.4. FINAL MIRDUPLEXSVM MODEL PREDICTIONS FOR EAE UP TO 5NTS.

		End Absolute Error (EAE) in nts					
		≤ 0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Prediction Accuracy (%)	k55	55.46	64.66	71.33	75.62	79.44	82.25
	k53	39.4	59.04	70.88	76.14	80.24	82.97
	k35	54	64.85	71.38	76.42	80.1	82.87
	k33	43.35	61.53	70.18	75.58	80.1	82.87

The absolute error for each one of the four ends of the duplex is calculated independently.

We next tested our methodology on the problem of identifying the mature molecule that lies on the opposite strand of a known miRNA. This is an important problem as both of these molecules are frequently functional, albeit under different conditions, and thus experimental techniques are unlikely to detect them both in a single experiment. To the best of our knowledge, this is the first attempt to find opposite strand miRNAs using a machine learning approach.

Towards this goal, we set the known miRNA of each hairpin as the ground truth for that strand and produce all candidate duplexes generated by sliding along the opposite strand. The final prediction is the highest scoring candidate. MiRduplexSVM is compared to a simple classifier, termed "Overhangs Ruler", which uses the statistical distributions of overhang lengths in the training set to identify the most frequently occurring values for 3' and 5' strands. In the majority of the cases, these values are equal to 2nts, a number that is commonly used in computational studies to find the miRNA* [57], [55]. For a new test hairpin, the missing strand of the duplex is estimated by assigning the overhang lengths to the known miRNA ends.

Both algorithms were trained on a dataset of 3.248 hairpins (containing a known duplex) and evaluated on a set of 2.000 hairpins (with known duplexes) using the EAE metric. Prediction accuracies were measured for each strand independently and the results are listed in Table 2.5. MiRduplexSVM was found to outperform the Overhangs Ruler on identifying the start position of the

miRNA* (Table 2.5, rows 1 and 3), while both algorithms achieve the same performance on predicting the end position (Table 2.5, rows 2 and 4).

TABLE 2.5. MISSING DUPLEXES PREDICTION RESULTS FOR MIRDUPLEXSVM AND THE OVERHANGS RULER.

End Absolute Error (EAE) in nts				
Prediction Accuracy (%) of MiRduplexSVM / Overhangs Ruler		≤ 0	≤ 1	≤ 2
	k55	70 / 56 ***	85 / 84 ns	91 / 92 ns
	K53	53 / 53 ns	79 / 81 ns	90 / 91 ns
	k35	67 / 53 ***	85 / 81 ***	91 / 91 ns
	k33	58 / 56 ns	82 / 84 ns	89 / 92 ns

*MiRduplexSVM outperforms Overhangs ruler in the identification of the start position of the mature miRNA that lie on the 5 or the 3' strand, but achieves the same accuracy on the prediction of their end positions. Statistical significance was assessed using the Fisher exact test. *** corresponds to p-value ≤ 0.001, ** to p-value ≤ 0.01, * to p-value ≤ 0.05 and ns to non significant.*

Our predictions were also contrasted to the results of a comparative genomics approach, a method frequently employed to find conserved miRNAs [92, 93]. Opposite strand molecules were identified by searching for orthologs in other species, utilizing the gene name of each miRNA. Orthologs with known duplexes were used to predict opposite strand miRNAs as long as (a) the known miRNAs were exactly the same across species and (b) the sequence of the opposite strand molecule was part of the hairpin under investigation. It is important to mention that if more than one orthologs met these requirements, several predictions were produced per hairpin. In this case, only the prediction with minimum EAE was used for comparison and thus the results provide an upper bound of the performance using orthologs based on best-case analysis. This process resulted in the identification of opposite strand miRNAs for 30 genes, while we note that

the MiRduplexSVM is capable of providing predictions every time. When compared to the MiRduplexSVM predictions for the same hairpins using the EAE metric, both methods gave the same predictions within a window of 2nts deviation (Table 2.6).

TABLE 2.6. MIRDUPLEXSVM VERSUS COMPARATIVE GENOMICS ON MISSING DUPLEXES PREDICTION.

End Absolute Error (EAE) in nts				
Prediction Accuracy (%) of MiRduplexSVM with respect to the comparative genomics results		≤ 0	≤ 1	≤ 2
	k55	81.82	95.45	100
	K53	63.64	95.45	100
	k35	85.71	100	100
	k33	100	100	100

The table shows the prediction accuracy per corner of MiRDuplexSVM when the results of a comparative genomics approach are set as the ground truth. When considering an error tolerance of up to 2nts, MiRduplexSVM gives exactly the same predictions as a strict comparative genomics algorithm.

Finally, we used MiRduplexSVM to predict all missing duplexes of human and mouse hairpins (1240 mature miRNAs, see Appendix).

Microprocessor cleavage site determination

As mentioned earlier, there are two biological models on how the microprocessor complex recognizes and process a pri-miRNA. The first model suggests that the Drosha cut site is located at ~22 nucleotides from the terminal loop – stem junction [6] while the second model claims that the cleavage site is located at ~11nts from the stem – single stranded tails junction [3]. We use our

algorithm, MiRduplexSVM to investigate concordance with these two hypotheses by performing *in silico* mutagenesis experiments. It is important to mention that, these junctions are not easy to define based on the secondary structure of miRNA hairpins. It has been shown experimentally that 8 out of 10 times, the miRNA hairpin's secondary structure, and especially their loop size and actual folding, is different from its computational prediction as shown by chemical and enzymatic probing [94]. Taking into consideration these inconsistencies and in accordance with Han et.al[3], we define two main regions on a given hairpin: region L, which includes 13 (upstream) and 11 (downstream) nucleotides from the Drosha site, and region U which includes all nucleotides between the Drosha cleavage site and the terminal loop tip, as shown in Figure 2.8. Our approach relies on the tip of the loop, and not its starting position, in order to avoid errors that may have been introduced during the secondary structure generation as discussed above.

A hairpin consists of a double-stranded part, the stem, and a sequence of unmatched nucleotides that connects the strands of the stem, called the terminal loop. The strand before the terminal loop is called the 5' arm of the hairpin while the other is called the 3' arm. The arms are not perfectly complementary but they form small loops and bulges. A miRNA:miRNA* duplex consists of two hairpin subsequences on each of the two arms, called the *5' strand* and the *3' strand* of the duplex. We can define a duplex by the positions of its four ends on the

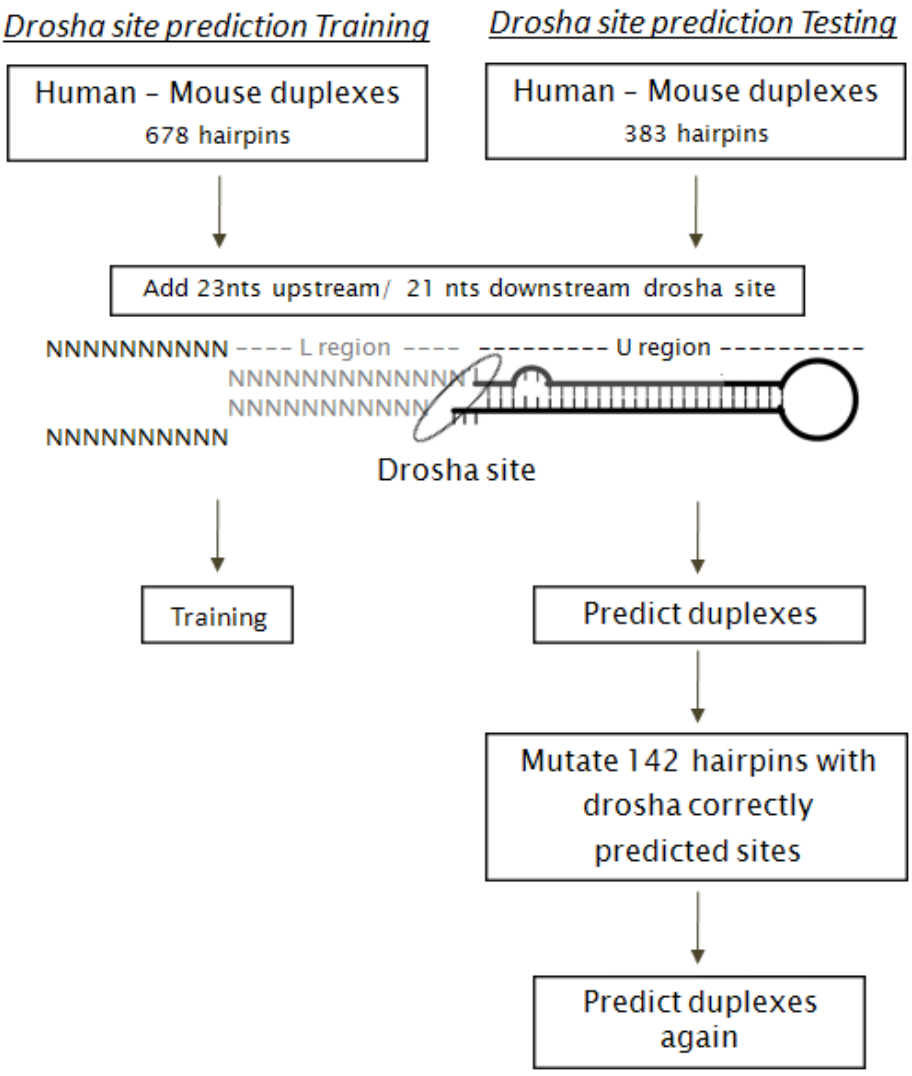

Figure 2.8. Flowchart of the in silico mutagenesis process. The L region corresponds to 13nts before and 11 nts after the Drosha cut site. The U region corresponds to the hairpin starting at the Drosha site and containing the loop.

generating hairpin sequence; we name them *k55*, *k53*, *k35* and *k33* corresponding to the *5'strand 5'end*, *5'strand 3'end*, *3'strand 5'end* and *3'strand 3'end* positions, respectively. Notice that, because of the way of counting positions, k55 < k53 < k35 < k33 [88].

To calculate the tip we identify the last matching nucleotides before the tip, which correspond to the loop start and loop end position, respectively. If the tip is T and the last matching nucleotides are X for 5' strand and X' for 3' strand, then X < T < X' and T = X + ceil ((X'-X) / 2), ceil refers to rounding toward positive infinity.

In addition, *prediction error* is assessed using Drosha Corner Sum Absolute Error, DCSAE. The DCSAE is the sum of absolute errors in number of nucleotides from true position between the actual and the predicted Drosha site end, taken over both ends of the Drosha site. For example, if the true positions of Drosha site are k55 = XX55 and k33 = XX33 and the predicted positions are YY55, and YY33 respectively, then the DCSAE = $|XX_{55} - YY_{55}| + |XX_{33} - YY_{33}|$.

In order to characterize the effect of nucleotide mutations on Drosha processing we used all human and mouse hairpins in miRBase 19.0 as graphically illustrated in Figure 2.8. To ensure the presence of the stem – single stranded tails junction in our sequences and the sequence - structure around it, we added whenever needed, 23 (upstream) and 21 (downstream) nucleotides from the Drosha site [3]. Out of this dataset, MiRduplexSVM was trained with the same hairpins that were used during parameter optimization [88] (678) and the remaining 383 hairpins were used for testing.

Only 142/383 hairpins whose Drosha sites were predicted correctly (0nts deviation) were used for the mutagenesis experiments. *In silico* mutagenesis was performed by inserting or deleting 2, 4 or 6 matching nucleotides to L or U regions, thus shifting the Drosha site towards or away from the single stranded tails – stem junction, or the terminal loop tip. MiRduplexSVM was re-applied on

the mutated hairpins and the new Drosha processing sites were predicted, Figure 2.8.

In order to evaluate if a mutation has a statistical significant effect on the MiRduplexSVM's predictions we perform Wilcoxon rank-sum tests between the results obtain using the Drosha Corners Sum Absolute Error (DCSAE) before and after every mutation. Bonferroni correction was also applied to correct for multiple testing.

Table 2.7 shows the statistically significant results from this comparison. The Drosha Site Average Shift (DSAS) and the percentage of hairpins in which a shift was observed after each mutation are also shown. In order to calculate the DSAS, only hairpins whose Drosha site had changed after a given mutation were taken into account. For each hairpin we first calculate Drosha Corners Mean Error, DCME = ((YY55 – XX55) + (XX33 – YY33))/2. Subsequently we estimate the median value over all hairpins' DCMEs, DSAS = median(DCMEs). As evident from the table, all mutations in the U region (between the Drosha site and the loop tip) resulted in a predicted shift in the Drosha site while for the L region (between the Drosha site and the stem – single stranded tails junction), only the deletion of 4nts led to significant changes.

TABLE 2.7. MUTATION ANALYSIS

Type of mutations	DSAS	Percentage %	Significance
L-4	-0,5	9.75	***
U+2	0.5	15.44	***
U+4	2.5	19.51	***
U+6	5	20.32	***
U–2	-1.5	20.32	***
U–4	- 4.5	43.9	***
U–6	-6.5	78.86	***

*In silico mutagenesis experiments. The first column lists the type of the performed mutations, i.e. the number of matching nts added (+) or deleted (-) in the L or U regions. The sequence of the inserted nucleotides was generated randomly. The second column shows the Drosha Site Average Shift (DSAS) in nts which corresponds to the median of the Drosha Corners Mean Error. The third column reports the percentage of hairpins in which a shift was predicted. The last column shows the statistical significance of these effects, assessed using a. Wilcoxon rank-sum test between DCSAE calculated on wild type and mutated sequences for each mutation. *** corresponds to p-value< 0.001*

Specifically, the deletion of 4 nucleotides in the L region resulted in a 0.5nt shift of the Drosha site towards the stem – single stranded tail junction (see Table 2.7). This finding is in contrast with recent experimental work whereby the deletion of 4 matching nucleotides in the respective L region of mir-16-1, pushed the Drosha site away from the stem – single stranded tail junction by the same distance[3]. A thorough analysis of our results, revealed that for an approximate 60% of the cases the Drosha cleavage site moves closer to the stem – single stranded tail junction by 1 nucleotide and the remaining 40% of the times it moves away by 2 nucleotides (see Table 2.8). Deletion of nucleotides form the U region however results in a shift of the Drosha cleavage site which is analogous to the direction of the mutation: inserting or deleting nucleotides moves the

Drosha site in a way that maintains specific distance from the stem loop tip. These findings are in close agreement with the experimental work of Yan Zeng et al [6].

TABLE 2.8. L REGION MUTATIONS

Type of mutations	DSAS	Percentage %
L-4	-1	58.3
L-4	2	41.6

The effect of L-4 type of mutations is shown. In 58% of the cases presenting an effect from this mutation the Drosha site moved one nucleotide from its original place towards the single stranded stem junction. In the remaining 42% of the cases, it moved 2 nucleotides towards to the stem loop junction.

CONCLUDING REMARKS

We introduced the problem of predicting the miRNA:miRNA* duplex stemming from a miRNA hairpin precursor as a first step in identifying the mature miRNA(s); the latter is important both for experimentally verifying the miRNA and for computationally predicting target mRNAs. We employed biological knowledge and constraints in converting the problem to a classification one and trained a high-order polynomial SVM model to identify the true duplex among candidates.

MiRduplexSVM outperforms existing approaches

We compared our methodology, named MiRduplexSVM, with (a) a distance based Simple Geometric Locator, and (b) three state of the art miRNA mature prediction tools, namely MatureBayes[55], MiRPara [56] MaturePred[57] and MiRdup[58]. In all cases, comparisons were performed in a fair and unbiased manner, employing the training set of each respective tool and evaluating performances on a common hold out test set. We showed that (a) for mammalian hairpins, MiRduplexSVM greatly outperforms all other tools in identifying either the true duplex or each of its ends independently. (b) For plant hairpins, our tool outperformed MaturePred_Plants on duplex prediction and matched its performance on finding each end of the miRNA molecules independently. The latter maybe due to the parameter optimization of MiRduplexSVM which was

done using mammalian hairpins and/or the small number of plant hairpins used to train MiRduplexSVM (198) compared to MaturePred_Plants (1.323). This finding is likely to change if a new version of MiRduplexSVM devoted to plant miRNA prediction was developed. Overall, that large improvement in performance seen especially for the 0nt deviation point (i.e. identification of the exact miRNA duplex or molecule) suggest that MiRduplexSVM is a much more efficient tool for addressing the problem of mature miRNA identification than currently available approaches.

Features of MiRduplexSVM that may underlie its high performance

The reasons behind this remarkable increase in accuracy achieved by MiRduplexSVM are multiple: first, our tool is trained to recognize miRNA:miRNA* duplexes, as opposed to strand-specific miRNAs, which is the standard approach of existing tools. Duplex formation is an indispensable stage in the biogenesis of all miRNAs, regardless of which strand will end up producing the functional molecule [8]. MiRduplexSVM takes into account this biological process and while it does not learn to distinguish which of the two strands is the functional miRNA, our results show that learning duplexes is also a very successful strategy for identifying strand specific miRNAs. Second, our tool learns to identify both the start and the end positions of the miRNA:miRNA* sequences, while most existing tools predict only the starting nucleotide and use a fixed size length of 22nts to find the end position. To achieve this

MiRduplexSVM uses a variable length parameter for each miRNA molecule. As a result, MiRduplexSVM does not only outperform other tools in predicting the start position of strand-specific miRNA molecules, but it also succeeds in specifying their length. Third, MiRduplexSVM does not assume a fixed size (2nt) overhang length like most existing approaches. On the contrary, the length of each overhang is explicitly learned by the training examples. This feature is likely to also contribute to the algorithm's success.

The zero padding, flank regions, and the representation to a fixed-vector size may also be a performance factor. The use of different cost hyper-parameters in the objective function of the SVM to handle the imbalance of the positive and negative classes is also a key factor. The candidate duplex generation also required the design of a simple, yet important algorithm to compute the overhang of the generated candidate. Finally, our experience with training the models shows that optimization and tuning of the SVM hyper-parameters (cost and polynomial kernel degree) is crucial for achieving good performance.

An important advantage of our model is its simplicity and cost effectiveness that results from the use of sequence information alone, as opposed to structure and thermodynamics that are often used by other tools [55-57]. In our tool, the secondary structure of hairpins is used in the preparatory stage, for filtering out multibranch precursors and estimating the distribution of overhang and mature miRNA lengths, but not for training the classifier. Actually, the incorporation of structural features in the model was not found to significantly affect performance (data not shown) and therefore was not explored further. Our results suggest that the information needed to identify miRNA:miRNA* duplexes

lies in the nucleotide sequence and MiRduplexSVM is capable of decoding it, making it a more accurate tool and at the same time less complex/expensive than others.

A tool for finding the complementary part of known miRNAs

We show that MiRduplexSVM is highly successful in identifying the complementary mature molecule (miRNA or miRNA*) that lies on the opposite strand of a known miRNA. It performs much better than the commonly used approach, where a fixed 2nt overhang is assumed in order to find the starting position of the opposite strand molecule. Moreover, it provides information for both the start and end positions, and therefore the length, of the unknown miRNA molecule. Finally, for human and mouse hairpins, we show that MiRduplexSVM behaves like a comparative genomics algorithm with strict criteria on this particular task.

A tool for performing *in silico* mutagenesis experiments

In addition to its high performance on mature miRNA identification, MiRduplexSVM was used to explore the effect of mutations on determining the Drosha cleavage site. There are currently two biological models regarding the

determination of the Drosha cleavage site: According to the model of Han et al [3], the complex cuts ~11nts from the stem – single stranded tails junction. According to the model of Zeng et al [6, 95], the microprocessor complex recognizes and cleaves a pri-miRNA ~22nts from the stem – loop junction. Our results suggest a third model that combines information from both of these studies.

In silico addition and/or deletion of matching nucleotides showed that the region *before* (U region in Figure 2.8) is more important than the region *after* (L region in Figure 2.8) the Drosha processing site. Specifically, every mutation in the U region resulted in a shift of the Drosha cleavage site while only the deletion of 4 in the L region had a similar effect. Our findings are in agreement with the recent work of Vincent et al. [96], where the Microprocessor complex was shown to distinguish between hairpins from different species by relying on sequence motifs that lie either within the single stranded tail region (U region) or the loop region (L region). These motifs have been suggested to guide the Microprocessor complex towards its cleavage site. It is possible that the insertion/deletion of nucleotides in the L and U regions that we simulate with MiRduplexSVM alters either the motifs themselves or the distances between the motifs and the Drosha cleavage site, thus resulting in the observed shift in the cleavage site itself. In sum, both ours and previous experimental findings suggest that structural and sequence information on both sides influence the Drosha cleavage point.

CHAPTER III – SIMPLE GEOMETRIC LOCATOR

INTRODUCTION

Supervised machine-learning approaches are frequently applied on biological data to learn a regression or classification model, whether used for prediction, classification, or for gaining an understanding on the biological process that has generated the data [97]. Arguably however, it is sometimes the case that sophisticated and complicated methods are employed, published, and advocated as advances without a comparison even against the simplest baseline methods. We consider a baseline method as the simplest method that an expert analyst can conceive within a few minutes of consideration of the problem and does not require any engineering or scientific ingenuity or novelty. A baseline can take the form of comparing against predicting by the mean of the outcome on a known dataset, without use of any special predicting variables, or comparing against random guessing. Because of this lack of comparison, the added-value of the sophisticated methods – if any – is not quantified; it remains unknown whether the extra effort for implementing or applying it is worth. A false perception about the difficulty of the problem may be created.

We now present an example of the above argument on the problem of identifying the position of miRNA mature molecules on their precursor RNA molecules, which typically have a hairpin-like secondary structure. In the cell, the miRNA precursor is first cut into a complex of two substring sequences (strands) with

high complementarity called the *5' strand* and the *3' strand*. The complex is called the miRNA:miRNA* duplex defined by its four corners denoted as k_{55}, k_{53}, k_{35} and k_{33} corresponding to the 5'strand 5'end, 5'strand 3'end, 3'strand 5'end and 3'strand 3'end positions, respectively (Figure. 1). The two strands are then separated and either one or both become a functional miRNA. The task is to predict the positions k_{55}, k_{53}, k_{35} and k_{33} given the sequence of a miRNA precursor molecule. Solving the problem can suggest novel miRNAs within suspected miRNA precursor sequences, to guide miRNA discovery, as well as provide intuition regarding the mechanisms regulating the miRNA biogenesis.

METHODOLOGY

The first paper to address this prediction task is ProMiR[50] which did not use any baseline method. Subsequent methods used the previous methods as baselines without considering the simplest possible method. What is the baseline "straw man" on this problem? Arguably, it is making a prediction based on the mean position of each corner k_{55}, k_{53}, k_{35} and k_{33} as estimated from a training set of known miRNA duplexes and their precursor sequences. We call this method the Simple Geometric Locator (SGL) obviously providing a constant predicted position on any hairpin independent of the input sequence. An important detail to address is to define the reference point for measuring the mean position since miRNA precursors have various lengths. We chose the terminal loop tip as the reference point as it does not depend on the length of the pre-miRNA flanking regions included in the hairpin sequence (details are in Chapter II).

The set of "cannons" to compare against SGL form four of the state-of-the-art tools for the task, namely MatureBayes[1] [55], MiRPara [56], MaturePred [57] and the most recent MiRdup [58] published in respectable venues such as PLoS ONE, BMC Bioinformatics, and Nucleic Acids Research. These tools employ machine-learning algorithms such as the Simple-Bayes Classifier, Support Vector Machines, and the random forest classifier. They also employ complex raw and constructed features that include the nucleotide sequence, the secondary

[1] We note that MatureBayes did compare against the SGL in Gkirtzou, K. (2009) Mature MiRNA Identification via the use of naive Bayes classifier. University of Crete, Computer Science Department.; unfortunately, the reference point used for the SGL was the beginning of the flanking regions, whose length is arbitrarily chosen before miRNA precursors are inserted in the MiRBase; hence, the performance of the SGL was found inferior in that work.

structure, number of loops and bulges, matches or mismatches for each nucleotide and others.

In our comparison, *prediction error* on the task for each corner (end) is measured as the End Absolute Error (EAE): the absolute error of the predicted minus the true position (in nucleotides) for a specific duplex end. See supplementary file 1 for an example. To measure *prediction accuracy*, we define as "correct" a prediction with error less or equal to a number x, i.e., EAE $\leq x$. Then, the prediction accuracy for an error bound (tolerance) of at most x, denoted as Accu(x), is the percentage of correct predictions in the test set. For example, if a model identifies the position of a given duplex end in 50% of duplexes within at most ±4nt from their true position, it has accuracy at 4nt of 50%: Accu(4) = 0.5 . *Statistical significance* of the results is assessed by assuming the null hypothesis that two methods have the same accuracy for a given error bound and applying the Fisher's exact test. To ensure fairness, in each comparison the SGL is trained (estimates the mean positions) with each method's training set, the one employed in the corresponding publication, after removing all miRNA hairpins with unknown duplexes and multi-branch structures (structures with more than one stems are considered multi-branch). Other programs' predictions are obtained and summarized in Chapter II. The performance accuracies are estimated on a common hold-out test set, as detailed in Supplementary file 1. Since some tools do not provide a prediction on all hairpins, the estimation of accuracy is computed only on the hairpins for which a prediction is made; SGL of course, is always providing with a prediction.

RESULTS

The results are shown in figure 3.1 and table 3.1. First, the accuracies $Accu_i(x)$ are computed for each tool i and duplex end for an EAE of x = 0,1,2, ... 5nt. Subsequently these accuracies are plotted against each other by connecting the points $(Accu_i(0), Accu_{SGL}(0)), (Accu_i(1), Accu_{SGL}(1)), \ldots, (Accu_i(5), Accu_{SGL}(5))$. For example, a point $(Accu_i(1)= 30\%, Accu_{SGL}(1)= 40\%)$ implies that method i identified 30% of the duplexes in the test set within ±1nt of their true position, while SGL identified 40% of duplexes within ±1nt of their true position. Thus, if a line is on the diagonal, then the two methods achieve the same accuracy for the same error tolerance. If it is below the diagonal SGL achieves lower accuracy for the same error tolerance and if it is above the diagonal, then the SGL achieves higher accuracy for the same error tolerance than the method compared against.

SGL clearly and statistically significantly outperforms MatureBayes and MaturePred in predicting any of the four duplex corners for all error bounds. On the 5' strand SGL and MiRdup achieve similar accuracies for absolute error of at most 0nt and 1nt with MiRdup slightly improving for larger error bounds; MiRPara on the 5' strand is better by only 2% - 6%. In the 3' strand, MiRdup and MiRPara exhibit an overall better performance than SGL. However, when focusing on the accuracy with zero tolerance $Accu(0)$, i.e., the percentage of duplexes identified on their exact position corresponding to the first point of each line, only MiRPara shows statistically significantly better results with the difference in performance ranging from 3%-10% (see Figure3.1 and table3.1).

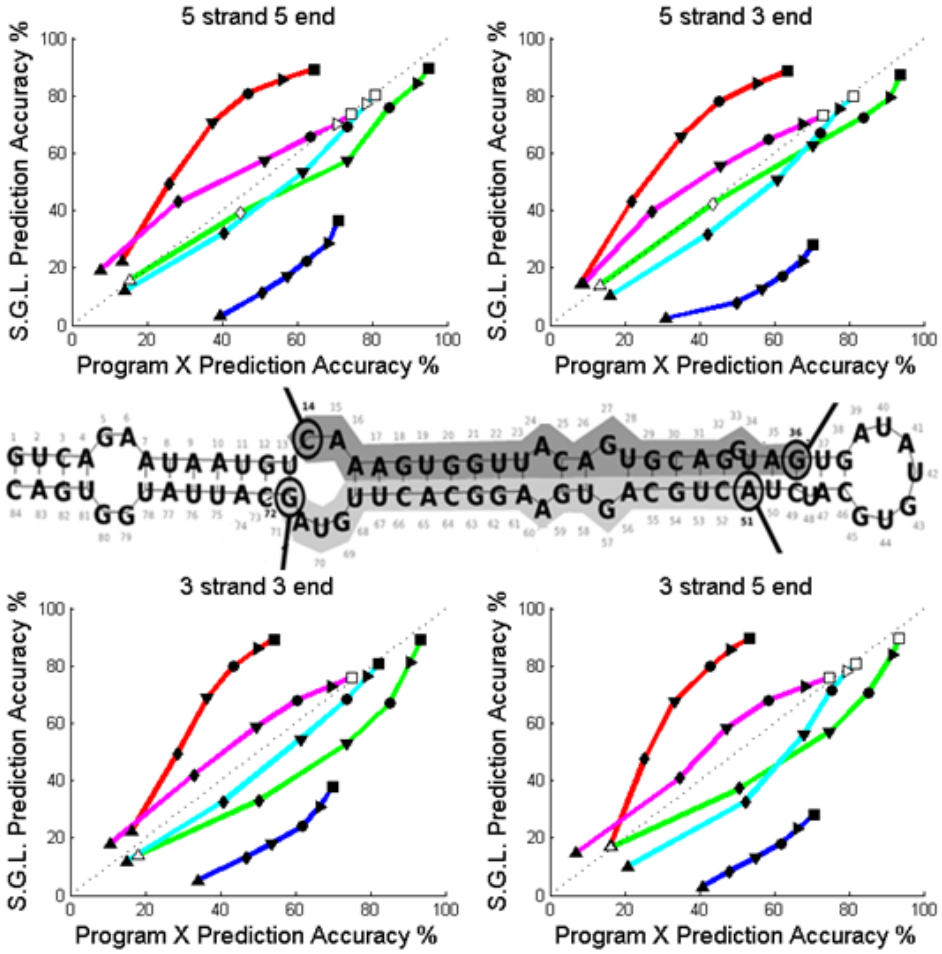

Figure 3.1. Prediction performance per corner. Performance accuracies are estimated using the EAE for up to 5nts. In each subplot, the y axis shows the prediction accuracy of Simple Geometric Locator (in %) and x axis shows the prediction accuracy of other methods (in %) for the same error tolerance. Lines comparing against the Simple Geometric Locator correspond to MatureBayes (magenda), MaturePred_Mammals (red), MaturePred_Plants (blue), MiRPara (cyan), and MiRdup (green) . In addition, upward triangle, diamond, downward triangle, circle, right pointed triangle, square, left pointed triangle, pentagram star and hexagram star correspond to errors less or equal than 0, 1, 2, 3, 4, 5 nucleotides, respectively. Statistically significant results are indicated with filled symbols.

TABLE 3.1 PREDICTION ACCURACIES, UP TO 8 NTS DEVIATION.

		End Absolute Error (EAE)								
		<= 0	<= 1	<= 2	<= 3	<= 4	<= 5	<= 6	<= 7	<= 8
Simple Geometric Locator/ MatureBayes - Prediction accuracy %	k55	19.44 %/ 7.79 % ***	42.94% / 28.34% ***	57.3%/ 51.19% ***	65.74% / 63.68% *	70.09% / 70.58% ns	73.59% / 74.48% ns	76.25% / 77.44% ns	78.04% / 79.61% ns	79.02% / 81.35% **
	k53	14.28 %/ 8.96 % ***	39.69% / 27.23% ***	55.4%/ 45.55% ***	64.69% / 58.44% ***	69.92% / 67.54% *	73.1%/ 73.13% ns	75.62% / 76.74% ns	77.52% / 79.72% *	78.72% / 81.27% **
	k35	14.81 %/ 6.9% ***	40.82% / 34.66% ***	58.33% / 47.2% ***	68.11% / 58.49% ***	72.79% / 68.4% ***	75.75% / 74.85% ns	77.34% / 78.85% ns	79.35% / 80.75% ns	80.65% / 82.63% *
	k33	17.98 %/ 10.66 % ***	41.64% / 32.89% ***	58.67% / 49.68% ***	68.09% / 60.47% ***	72.82% / 69.62% **	75.86% / 75.12% ns	77.87% / 79.16% ns	79.64% / 81.31% *	80.96% / 83.18% **
Simple Geometric Locator/ MiRPara - Prediction accuracy %	k55	12.01 %/ 14.03 % *	32.23% / 40.65% ***	53.25% / 61.68% ***	69.13% / 73.44% ***	77.06% / 78.59% ns	80.19% / 80.93% ns	81.9%/ 82.6% ns	83.4%/ 84.79% ns	85%/ 85.66% ns
	k53	10.2 %/ 16.05 % ***	31.85% / 42.15% ***	50.75% / 60.81% ***	66.97% / 72.26% ***	75.29% / 77.62% *	79.71% / 81.17% ns	82.04% / 82.81% ns	83.26% / 83.99% ns	84.44% / 85.03% ns
	k35	10.05 %/ 20.95 % ***	32.63% / 52.3% ***	56.06% / 67.98% ***	71.67% / 75.32% ***	78%/ 79.55% ns	80.81% / 81.72% ns	83.04% / 83.38% ns	84.43% / 84.77% ns	85.61% / 86.15% ns
	k33	11.65 %/ 14.93 % ***	32.7%/ 40.72% ***	54.4%/ 61.37% ***	68.25% / 73.73% ***	76.34% / 79.18% **	80.64% / 82.36% *	83.01% / 84.5% ns	84.83% / 85.78% ns	86.15% / 86.97% ns
Simple Geometric Locator/ MaturePred_Mammals - Prediction accuracy %	k55	22.33 %/ 13.44 % ***	49.17% / 25.94% ***	70.64% / 37.62% ***	80.81% / 46.94% ***	85.8%/ 56.1% ***	89.31% / 64.47% ***	90.97% / 69.69% ***	92.26% / 74.11% ***	93.35% / 80.33% ***
	k53	14.39 %/ 8.31 % ***	43.28% / 22% ***	65.75% / 35.11% ***	78%/ 45.13% ***	84.51% / 55.49% ***	88.65% / 63.66% ***	91.02% / 69.12% ***	92.4%/ 74.16% ***	93.82% / 80.1% ***
	k35	17.01 %/ 16.03 % ns	47.5%/ 25.3% ***	67.66% / 33.23% ***	79.8%/ 42.96% ***	85.59% / 48.33% ***	89.48% / 53.29% ***	91.15% / 57.78% ***	93.19% / 62.23% ***	94.25% / 66.91% ***
	k33	22.2 %/ 16.4 % ***	49.21% / 28.64% ***	68.86% / 36.61% ***	79.8%/ 43.56% ***	85.87% / 49.95% ***	89.34% / 54.36% ***	91.52% / 59.64% ***	92.96% / 64.32% ***	93.93% / 69.23% ***
Simple Geometric Locator/ MaturePred_Plants - Prediction accuracy	k55	3.2%/ 39.62 % ***	11.22% / 50.69% ***	17.19% / 57.47% ***	22.36% / 62.71% ***	28.4%/ 68.32% ***	36.42% / 71.01% ***	43.41% / 74.58% ***	49.02% / 76.84% ***	53.97% / 78.88% ***

%	k53	2.33 %/ 30.95 % ***	7.79%/ 49.96% ***	12.53% / 56.74% ***	17.04% / 62.2% ***	22.36% / 67.52% ***	28.11% / 70.21% ***	36.27% / 73.71% ***	42.39% / 77.2% ***	46.61% / 78.51% ***
	k35	2.95 %/ 40.68 % ***	8.4%/ 48.05% ***	13.19% / 55.2% ***	18.05% / 61.97% ***	23.07% / 66.4% ***	28%/ 70.52% ***	34.27% / 73.69% ***	39.2%/ 76.49% ***	45.17% / 78.78% ***
	k33	5.31 %/ 33.9 % ***	13.19% / 46.79% ***	18.05% / 53.87% ***	23.95% / 61.75% ***	30.58% / 66.54% ***	38.03% / 70.08% ***	43.26% / 73.25% ***	47.75% / 76.27% ***	52.25% / 79% ***
Simple Geometric Locator/ MiRdup - Prediction accuracy %	k55	15.86 %/ 15.52 % ns	39.31% / 44.83% ns	57.24% / 73.45% ***	75.86% / 84.48% **	84.14% / 92.07% **	89.66% / 95.17% **	92.76% / 96.21% ns	95.52% / 97.59% ns	97.24% / 97.59% ns
	k53	14.14 %/ 13.45 % ns	42.41% / 43.45% ns	62.76% / 70.34% *	72.41% / 83.79% ***	79.31% / 91.03% ***	87.59% / 93.79% **	92.07% / 95.52% ns	94.48% / 97.93% *	96.9%/ 97.93% ns
	k35	16.9 %/ 16.55 % ns	37.24% / 50.69% ***	56.9%/ 74.83% ***	70.69% / 85.17% ***	83.79% / 91.72% **	89.66% / 93.45% ns	92.41% / 94.83% ns	94.83% / 95.86% ns	95.52% / 97.24% ns
	k33	14.14 %/ 18.28 % ns	33.1%/ 50.34% ***	52.76% / 73.79% ***	66.9%/ 85.17% ***	81.38% / 90.69% ***	88.97% / 93.45% *	91.38% / 94.48% ns	94.83% / 96.21% ns	97.24% / 97.59% ns

*The absolute error for each one of the four ends of the duplex is calculated independently. Simple Geometric Locator has been trained on each program's training set and their performance has been accessed on a common blind test set. Fisher exact test were performed to examine if the observed differences are statistical significant. *** corresponds to pvalue ≤ 0.001, ** to pvalue ≤ 0.01, * to pvalue ≤ 0.05 and ns to non statistical.*

CONCLUDING REMARKS

Comparing against the simplest possible method as a baseline in data analysis is an important step of the analysis. Foregoing this step may result in unnecessary effort and energy spent in code developing, publishing, and evaluations by future researchers, unnecessary use of computationally expensive methods and a false impression about their benefits and added value they provide in a given task. As an example we show that using the mean positions in predicting the four corners of a miRNA duplex complex outperforms some state-of-the-art methods and is on par with the rest when trying to predict the exact location of the duplex with zero tolerance.

CHAPTER IV – IDENTIFICATION OF THE MATURE SEQUENCE OF FOUR miRNA CANDIDATES

In this part of the thesis we present the experimental identification of the mature sequence of recently identified miRNA candidates [5] that are located in a cancer associated genomic region frequently deleted in bladder cancer [93]. We named them c-miR-ch9, c-miR-ch5a, c-mir-ch5b and c-miR-ch22 and they are located at chr9:123327358-123327460 strand-, chr5:148958951-148959053 strand–, chr22:40863894-40863996 strand+ and chr5:149984684-149984786 strand- respectively [5]. In addition we predict cyclin D2 (CCND2), a gene with documented oncogenic activity, [98] as a key target of c-miR-ch9 and validate this interaction using luciferase reporter assays.

It is worth noting that, in addition to our scientific findings, this is the first, to the best of our knowledge, integrative approach in which the prediction of putative pre-miRNAs is followed by the experimental verification of their mature sequence and the computational prediction of a target is experimentally confirmed using reporter assays.

Mature miRNA prediction by primer extension

We designed three overlapping primers (each 15 nts in length) to bind to the verified positive strand of the each precursor sequence. As a positive control of the primer extension reaction hsa-let-7a-5p was selected and only one primer complement to its mature sequence was used. The primers were labeled using γ32 ATP and three primer extension reactions were performed under the following conditions: (A) incubation of 4 μg of Hela total RNA with the respective primer at 65°C for 5 min, followed by 1 min on ice; (B) subsequent incubation for 30 min at 16°C; (C) gradual increase in the temperature (0.1°C /sec) to 42°C and incubation for another 30 min at the later temperature. This gradual increase in temperature provides optimum conditions for primer extension and prevents the dehybridization of the primer. The reaction was terminated by incubation for 5 min at 85°C. In order to determine buffer's, dNTPs', reverse transcriptase's and RNase inhibitor's concentrations, we followed the HT SuperRTkit manufacturer's instructions.

Primers for c-miR-ch9:

- ✓ 5' ACC AGG GGA CAC CGT
- ✓ 5' CTG CCA GGT TCC ACC
- ✓ 5' TTA CCT CTC CCC CTG

Primers for c-miR-ch5a:

- ✓ 5’ GAA GAC AGG TGT CAT
- ✓ 5’ CCC CAG GCC CCC GAA
- ✓ 5’ TAC GCC CAC AGC CCC

Primers for c-miR-ch22:

- ✓ 5’ CGA CCG CCC GCC TGC
- ✓ 5’ CGA GGA CAC GGC CGA
- ✓ 5’ GGT GAC CCG CGG CGA

Primers for c-miR-ch5b:

- ✓ 5’ GCC CCT TCC CAC CTG
- ✓ 5’ CGC GCG AGC CTC GCC
- ✓ 5’ GGG TGC GGG CAC CGC

Primer for hsa-let-7a-5p

- ✓ 5’ AAC TAT ACA ACC TAC

RNA extraction and northern blot analysis

Total RNA was extracted from HeLa cells grown in culture using Trizol. Eighty micrograms of total RNA were analyzed with DNA oligonucleotides probes and 30 μg of total RNA were analyzed using LNA oligonucleotides on a 15% denaturing polyacrylamide gel containing 7 M urea and transferred to Nytran N membrane (Schleicher and Schuell). Membranes were probed with standard DNA or LNA oligonucleotides. Several DNA oligonucleotides concerning all

miRNA candidates and one LNA oligonucleotide concerning only c-miR-ch9 were used. For c-miR-ch5a, two DNA probes were complement arm to the strand which produces the mature miRNA and one was complement arm to the opposite strand [5]. For c-miR-ch5b, one DNA probe was complement arm to the predicted mature sequence and another one was complement to the adjacent sequence. For c-mir-ch22, two DNA overlapping probes were used, due to the fact that primer extension analysis gave two possible mature sequences. For c-miR-ch9 one DNA probe was complement arm to the predicted mature sequence and the other was complement arm to the adjacent sequence, which was used as a negative control. The LNA probe was complement arm to the predicted mature sequence. Ten picomoles of each DNA oligonucleotide probe and two picomoles of the LNA oligonucleotide probe were end-labeled with [γ-32P] ATP by using T4 polynucleotide kinase. Prehybridization of the ðlters was performed in 7% SDS, 5 × SSC, 1× Denhardt's solution and 0.02 M Na2HPO4 pH 7.2. Hybridizations were performed in the same solution at 50°C after the addition of the radiolabeled DNA oligonucleotide and at 60°C after the addition of the radiolabeled LNA oligonucleotide. Following an overnight hybridization, the membranes were washed at 50°C and 60°C, for DNA and LNA probes respectively, in low stringency buffer [2 x SSC, 0.3% SDS] twice for 30 min. An extra washing step was performed for LNA probes using 1 × SSC, 0.3% SDS, for 15 min, at 60°C. For DNA probes, membranes were stripped by washing in a high stringency buffer (0.1 × SSC and 0.5% SDS) for 30 min at 80°C and reprobed with the negative polarity oligonucleotides.

DNA probes for c-miR-ch5a:

- Positive 1: ACA GCC CCC AGG CCC CCG AAG ACA GG
- Positive 2: AGG CCC CCG AAG ACA GGT GTC ATG GA
- Positive 3: TAC GCC CAC AGC CCC CAG GCC CCC GA
- Negative: GGG GAG CCA GCA GGG AGG ACA TAC GC

DNA probes for c-miR-ch5b:

- Positive: GGG TGC GGG CAC CGC GCG AGC CTC GC
- Negative: CCT TCC CAC CTG CGC TAT TCC CGG CG

DNA probes for c-miR-ch22:

- Positive1: CGG CGA GGA CAC GGC CGA CCG
- Positive2: GAT GGT GAC CCG CGG CGA GGA

Probes for c-miR-ch9:

- *DNA*
  - Positive: TAC CTC TCC CCC TGC CAG
  - Negative: ACC AGG GGA CAC CGT GTG
- *LNA*
  - Positive: TAC CTC TCC CCC TGC CAG

To generate reporter vectors bearing miRNA-binding sites, we used two mammalian vectors phRL-TK (Promega, Madison, US) and pGL4-10 carrying the Renilla luciferase gene (hRluc) and firefly luciferase gene (luc), respectively. Specific oligonucleotides having XbaI ends and containing binding sites (b.s.) in triple repeats for the predicted c-miR-ch9::CCND2 interaction, were generated (Metabion). The phRL-TK vector was used for normalization. The oligos were cloned into the pGL4-10 vector at the XbaI site downstream of the luc gene. For all reporter constructs, two types of cassettes were prepared and studied side by side: wild type (pGL4-10 + wt—Triplet) and carrying mutations (pGL4-10 + mut—Triplet). We further PCR amplified the actual b.s. from the 3'UTR of CCND2 including ~500bp flanking regions on either side of the b.s. Following PCR mutagenesis of this construct (~1000bp), we cloned both wt-3'UTR (pGL4-10 + wt–3'UTR) and mut-3'UTR (pGL4-10 + mut-3'UTR) into the PGL4-10 vector. The empty vector (pGL4-10) was utilized as a control to observe the effect of our miRNA on the construct per se. All constructs were verified by sequencing. Additionally, anti-c-miR-ch9 LNA (Exiqon, Berlin, Germany) was used to inhibit the expression of c-miR-ch9. The sequences used in our studies are listed in Figure 4.1. Positions of mutations in the mutated constructs are indicated in bold.

a
Wild type interaction:

```
5′ AGTGGGGGCCGAGTTGTTCCCCCAGCCTGCCAA CCND2
                          |||||||
              3′  ATGGAGAGGGGGACGGTC c-miR-ch9
```

Wild type: GGGGGCCGAGTTGTTCCCCCAGCCTGCCAA

Mutated interaction:

```
5′ AGTGGGGGCCGAGTTGTTCCCCCAGCGAGGGAA CCND2
                          |   |   |
              3′  ATGGAGAGGGGGACGGTC c-miR-ch9
```

Mutant: GGGCCGAGTTGTTCCCCCAGCGAGGGAA

b
P1-FORWARD: TCTAGAATGGTGACTGACCCTTGAGC
P2-REVERSE: TCTAGAATCCTGCTAGCAATGGGATG
Mutagenesis Left : CAGCGAGGGAAATTTTGATCCTTCCCCTCTTT
Mutagenesis Right: ATTTCCCTCGCTGGGGGAACAACTCG

Figure 4.1. Cloned sequences and primers. (a) Mutant seed regions designed for pGL4-10 + mut – Triplet and pGL4-10 + mut-3'UTR. (b) Primers for CCND2 3'UTR amplification and PCR mutagenesis for pGL4-10 + mut-3'UTR. (c) LNA sequence

Transfection assay

Human HeLa 229 cell lines (LGC Promochem, ATCC Number: CCL-2.1) were grown in Dulbecco's Modified Eagle's Medium (DMEM) at 37°C, in a humidified atmosphere of 5% CO2. The cells were transfected in the 24-well plates in serum-free DMEM by using Lipofectamine 2000 (Invitrogen) according to manufacturers' instructions. For each transfection experiment, 350 ng of appropriate reporter construct, 50ng of normalization vector and 400 ng of pBSK(+) as a carrier plasmid, were used in order to obtain optimal results. HeLa

cells were also transfected with the empty pGL4-10 vector which was used as a reference point. Cells were harvested 48h after transfection and assayed both for firefly and Renilla luciferase activity using Dual Luciferase Assay System (Promega). The luciferase activity was measured using Dual Luciferase Assay System (Promega) with a FB 12 Luminometer (Berthold Detection Systems). For the inhibition of endogenous c-miRCh9 miRNA in HeLa cells the transfection of anti-c-miR-ch9 LNA (Dharmacon) at varying concentrations ranging from 25–50 nM was performed, using Lipofectamine 2000 according to manufacturer's instructions. Final expression values form transfection assays reported here were calculated by averaging all repeats for the particular construct. Values for error bars were calculated using the following formula for estimating the standard error of the mean: $\sigma M = \sigma/\sqrt{N}$, where σ is the standard deviation of the original distribution and N is the sample size (the repetition number).

RESULTS

Experimental identification of the mature miRNA sequence for four novel miRNA candidates

Our goal was to extract the mature (functional) miRNA sequence from the four potential pre-miRNAs, c-mir-ch5a, c-mir-ch5b, c-mir-ch9, c-mir-ch22, and show that these small RNA molecules are expressed in various cell lines. Unfortunately, according to recently produced deep sequencing data from HeLa cells [41], no small RNA sequences are expressed from the genomic location where the pre-miRNAs were detected. As a result, no prior experimental evidence was available regarding the location and/or sequences of the mature miRNAs. To address this problem, we used an adjustment of the primer extension methodology for identifying the most probable miRNA mature sequence from a putative precursor. Specifically, instead of using one primer complement to the mature sequence, which in our case was unknown, we designed three different overlapping primers that are complementary to the positive strand of the precursor sequence, namely the strand producing a small RNA (see Figure 4.2a). Using these primers we performed three primer extension reactions. The length of the extended primers further defines the location of the mature sequence on the precursor. The primer may bind to the precursor and/or the mature sequence. We assume that binding of the primer to both the mature as well as the precursor sequence will result in competition of binding and this will be evident in the banding patterns resulting from the primer extension reactions.

c-mir-ch9

For the first primer extension reaction the longest expected product is 19nt and the results show a band in the vicinity of 19nts (Figure 4.2b column 1). This product could be the result of either the extension of the precursor or the extension of the mature. Hence, the way that this band was generated remains to be verified by investigating the products of the other 2 primer extension reactions. If this band is a direct result of the extension of the mature, then the other 2 primers extension reaction would be expected to produce sequences of length more than 31nts, resulting from the binding and extension of the primer to the precursor sequence. Clearly this is not the case.

The second primer extension reaction produced one band (Figure 4.2b column 2) which is potentially derived due to the extension of the precursor. For this primer extension reaction the longest expected product through precursor binding and elongation is 31nts and the observed band corresponds to this length.

The observed banding patterns of the third reaction (Figure 4.2b column 3) reveal two products, one at 43nts and one at 19nts. We believe that the long band, which could be visualized only after 4 days of exposure, is due to precursor elongation since the longest expected product through precursor binding and extension would be 43nts. We consider the short band, which was detectable after overnight exposure, as a result of the extension of the mature sequence and we use this as our reference point for the prediction of the c-miR-ch9 mature sequence. Based on that, we determine that the 5' start of the mature sequence is at the 26th nucleotide of the precursor's sequence. Having identified the 5' end

of our mature sequence, and knowing that our mature sequence is 18nts long, we are able to deduce the 3' end of the mature sequence (Figure 4.2c). According to the results from our primer extension methodology we predict that the mature sequence for the potential miRNA (c-miR-ch9) is 5' CUGGCAGGGGGAGAGGUA.

Due to the novelty of the methodology (primer extension has never been used before for the prediction of the miRNA mature sequence) and the fact that some of the bands were faint we decided to verify our prediction with a northern blot analysis. We carried out a northern blot analyses using a DNA probe complement arm to the mature sequence as predicted by our primer extension reaction (black arrows in Figure 4.2c) and we also used a negative control, a DNA probe complement arm to the adjacent sequence (green arrows in Figure 4.2c, data not shown). Additionally, to increase sensitivity and signal of the experiment we also used an LNA probe complementary to the mature sequence (Figure 4.2d).

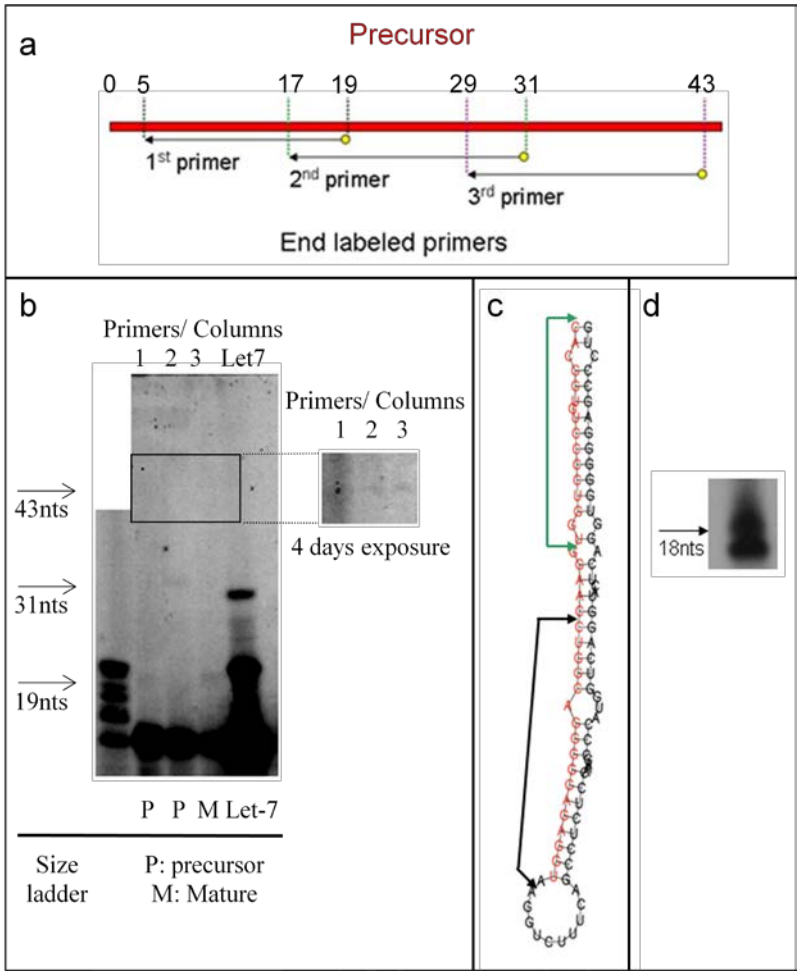

Figure 4.2. MiRNA mature prediction methodology, c-mir-ch9.
a) Three primers were designed to bind to the whole length of the verified positive strand of the hypothetical miRNA [5] c-miR-ch9. b) MiRNA mature prediction methodology results. Each column displays the banding patterns from the corresponding primers shown in (a). The banding patterns of the first and the second columns result from the annealing of the first and the second primers to the precursor (denoted by P) sequence. While the banding pattern on the third column is a consequence of the annealing of the third primer to the precursor, 43nts band, and the mature (M) sequence, 19nts band. c) Predicted potential mature miRNA sequence. Based on the results in (b) and the fact that the expected mature is 18nts long [5] we conclude that the mature sequence is between the 24th and the 41stnt on the verified positive (5p) strand. The black arrows indicate the predicted mature miRNA sequence. The green arrows indicate the sequence which was used as a negative control for the northern blot. d) Northern blot analysis. In order to verify the predicted mature miRNA sequence we perform a northern blot analysis using a DNA probe complement to the predicted mature, (black arrows in c), and as a negative control a DNA probe complement to the adjacent sequence, (green arrows in c). In order to increase our experiments sensitivity and signal an LNA probe complement to the predicted mature was also used. The upper observed band is ~18 nts long as expected from previous published data [5].

c-mir-ch5a

For the first primer extension reaction the longest expected product is 20nts and the results show a band in the vicinity of 20nts (Figure 4.3b column 1). This product could be the result of either the extension of the precursor or the extension of the mature. Hence, the exact binding remains to be verified by investigating the products of the other 2 primer extension reactions. If this band is a direct result of the extension of the mature, then the other 2 primers extension reaction would be expected to produce sequences of length more than 31nts, resulting from the binding and extension of the primer to the precursor sequence. Clearly this is not the case.

The second primer extension reaction produces one band (Figure 4.3b column 2) which is potentially derived due to the extension of the mature. For this primer extension reaction the longest expected product through precursor binding and elongation is 31nts but the observed band was 21nts.

No band was observed from the third reaction (Figure 4.3b column 3). We consider the short band (Figure 4.3b column 2), which was detectable after overnight exposure, as a result of the extension of the mature sequence and we used this as our reference for the prediction of the c-miR-ch5a mature sequence. Based on this, we determine that the 5' start of the mature sequence is at the 10^{th} nucleotide of the precursor's sequence. Having identified the 5' end of our mature sequence, and knowing that our mature sequence is 25nts long [5], we are able to deduce the 3' end of the mature sequence (Figure 4.3c). According to the results from our primer extension methodology we predict that the mature

sequence for the potential miRNA (c-miR-ch5a) is 5'CCUGUCUUCGGGGGCCUGGGGGCUGU.

Thereafter, we carried out a northern blot analyses using a DNA probe complement arm to the mature sequence as predicted by our primer extension reaction (black arrows in Figure 4.3b) and we also use a negative control, a DNA probe complement arm to the adjacent sequence (green arrows in Figure 4.3b). Northern blot analysis did not give the expected results. The "positive 3" (blue arrows in Figure 4.3b) and "negative" (green arrows in Figure 4.3b) probes produce the banding pattern shown in Figures 4.3ci and 4.3cii respectively. According to Oulas et.al [5] we expected the mature sequence to be 25nts long and to be produced from the 5' strand of the predicted precursor (Figure 4.3b, sequence highlighted in red). Surprisingly the band produced from that strand was 30 nts. In addition, the opposite strand also produced a band lower than 25nts. Because the reason for these inconsistencies was not clear, we decided to not devote more time and materials in the identification of the mature sequence of this specific precursor.

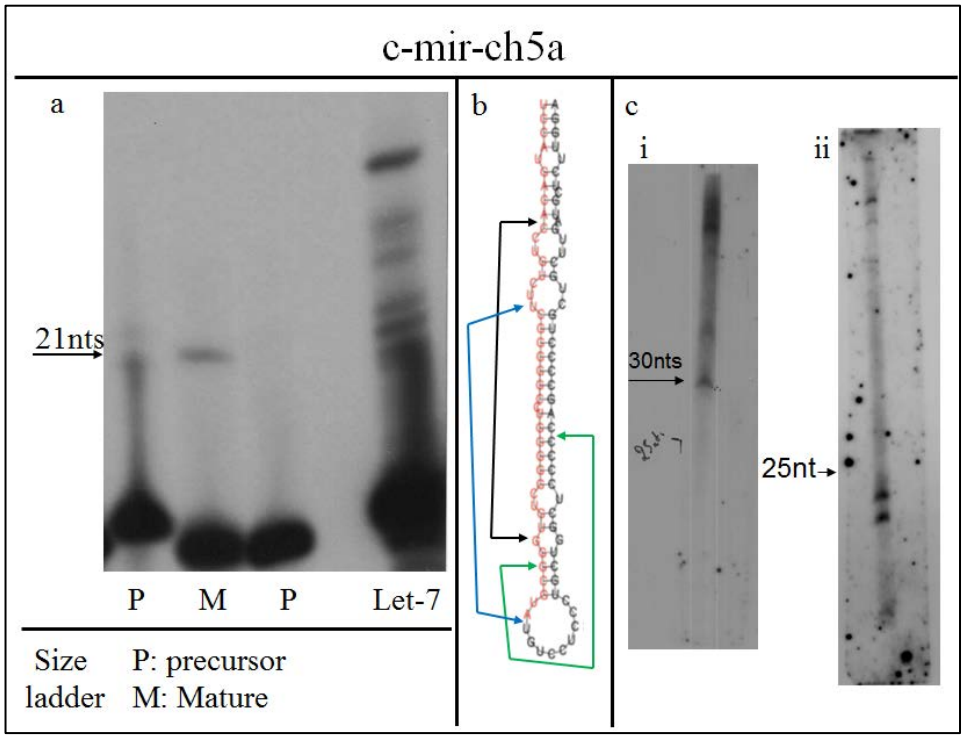

Figure 4.3. MiRNA mature prediction methodology, c-mir-ch5a.
a) MiRNA mature prediction methodology results. The banding patterns of the first column results from the annealing of the first primer to the precursor (denoted by P) sequence. While the banding pattern on the second column is a consequence of the annealing of the second primer to the mature, 21nts band, and the mature (M) sequence. b) Predicted potential mature miRNA sequence. Based on the results in (a) and the fact that the expected mature is 25nts long [5] we conclude that the mature sequence is between the 10th and the 36thnt on the verified positive (5p) strand. The black arrows indicate the predicted mature miRNA sequence. Blue arrows indicate the sequence which was randomly selected to investigate if a part of the precursor, other than the predicted mature, could produce a banding pattern. The green arrows indicate the sequence which was used as a negative control for the northern blot. c) Northern blot analysis. In order to verify the predicted mature miRNA sequence we perform a northern blot analysis using DNA probes complement to the predicted mature, (black - blue arrows in b), and as a negative control a DNA probe complement to the adjacent sequence, (green arrows in b). Note that the northern blot for the predicted mature sequence did not produce a band, (data not shown). i) The observed lower band – corresponding to the blue arrows – is ~30 nts, 5nts longer than expected from previous published data [5]. ii) The observed bands – corresponding to the green arrows – are less than 25nts long, and are produced from the 3' strand of c-mir-ch5a which was unexpected from previous published data[5].

c-mir-ch5b

For the first primer extension reaction unexpectedly no band was observed, (Figure 4.4a column 1). The second primer extension reaction produced two bands (Figure 4.4a column 2) which are potentially produced due to the extension of the precursor. For this primer extension reaction the longest expected product through precursor binding and elongation up to the start of the terminal loop is 31nts and the longest observed band corresponds to this length.

The observed banding patterns of the third reaction (Figure 4.4a column 3) revealed two products, one at 22nts and one lower than 27nts. We believe that the long band is due to precursor elongation and the short band (22nts) due to the extension of the mature sequence and we use this as our reference point for the prediction of the c-mir-ch5b mature sequence. Based on that, we determined that the 5' start of the mature sequence is at the 21th nucleotide of the precursor's sequence. Having identified the 5' end of our mature sequence, and knowing that our mature sequence is 25nts long [5], we were able to deduce the 3' end of the mature sequence (Figure 4.4b). According to the results from our primer extension methodology we predict that the mature sequence for the potential miRNA (c-miR-ch5b) is 5'GCGAGGCUCGCGCGGUGCCCGCACC.

Following the same methodology as in the previous examples, we carried out a northern blot analyses using a DNA probe complement to the mature sequence as predicted by our primer extension reaction (black arrows in Figure 4.4b) and we also used a negative control, a DNA probe complement arm to the adjacent sequence (green arrows in Figure 4.4b). The complement arm to the mature sequence DNA probe produced two bands (Figure 4.4c), one faint at 16nts and

one clear at 18nts. Because the miRNA mature sequences are between 18 – 27 nts long and the 18nts band was sharper than the 16nts band, we conclude that the mature sequence produces the longer band and perhaps the smaller band is due to degradation products. Based on both primer extension's reaction data and our Northern blot analysis we conclude that the mature sequence is: 5'GCGAGGCUCGCGCGGUGC – 18nts long.

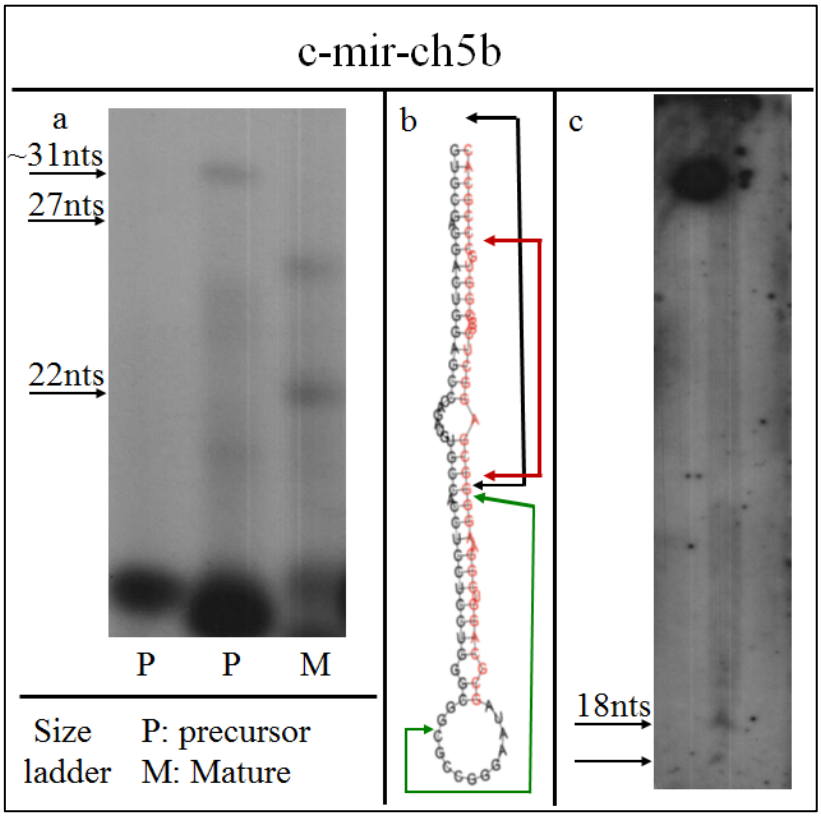

Figure 4.4. MiRNA mature prediction methodology, c-mir-ch5b.
a) MiRNA mature prediction methodology results. The first primer extension reaction, first column, did not result any bands. The banding pattern on the second column is a consequence of the annealing of the second primer to the precursor (denoted by P) sequence. While the banding pattern on the third column results from the annealing of the third primer to the mature sequence (M). The top band corresponds to the extension of the precursor and the small band (22nts) to the extension of the mature. b) Predicted potential mature miRNA sequence. Based on the results in (a) and the fact that the expected mature is 25nts long [5] we anticipate the mature sequence to be between the 21th and the 45thnt on the verified positive (3p) strand. The black arrows indicate the predicted mature miRNA sequence. The green arrows indicate the sequence which was used as a negative control for the northern blot. c) Northern blot analysis. In order to verify the predicted mature miRNA sequence we perform a northern blot analysis using DNA probes complement to the predicted mature, (black - arrows in b), and as a negative control a DNA probe complement to the adjacent sequence, (green arrows in b). The observed band is 18nts long. Based on the primer extension and northern blot's results, we conclude that the mature sequence is 18nts long and its sequence begins at the 21st nt and ends at the 38th nt (red arrows in b) on the verified positive 3p strand.

c-mir-ch22

The first primer extension reaction produced several bands (Figure 4.5a column 1). Their sizes range between 15nts to 27nts. The sharper band is 18nts long. This product could be the result of either the extension of the precursor or the extension of the mature. Hence, the exact binding remained to be verified by investigating the products of the other 2 primer extension reactions. If this band has a direct result of the extension of the mature, then the other 2 primers extension reaction would be expected to produce sequences of length more than 31nts, resulting from the binding and extension of the primer to the precursor sequence. Clearly this was not the case.

The second primer extension reaction (Figure 4.5a column 2) produced a similar banding pattern with the first reaction: a sharp band at 18nts and faint bands between 27 and 18 nts. The sharp band could be a result of either the extension of the precursor or the extension of the mature.

The observed banding patterns of the third reaction (Figure 4.5a column 3) reveal one product, at 18nts. The absence of other bands was probably due to competition of binding of the primer between the precursor and the mature sequence.

From the results of the primer extension reactions, we could not conclude which one of the two 18nts sharp bands produced by the 2nd and 3rd reaction, was due to the extension of the mature sequence. In addition because the expected length of the mature sequence was unknown[5] we set the length of the possible mature at 21nts.

Considering all the above, we predicted two possible miRNA mature molecules with the following sequences, 5' CGGUCGGCCGUGUCCUCGCCG (Figure 4.5b black arrows) and 5' UCCUCGCCGCGGGUCACCAUC (Figure 4.5b red arrows). To resolve which one of the predicted matures corresponds to the real one we perform northern blot analysis using probes complement to each one of them. Unfortunately, none of the Northern blots revealed any band (data not shown). Thus, we could not conclude the mature sequence of the c-mir-ch22 precursor.

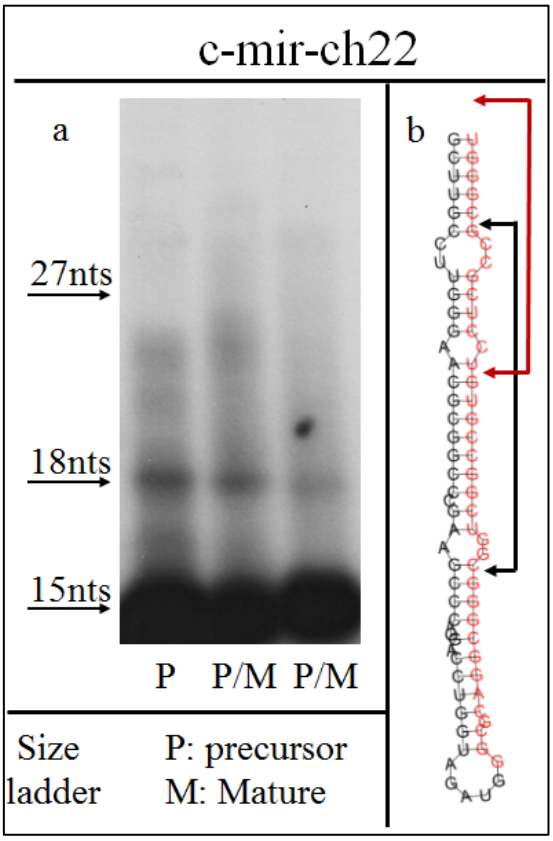

Figure 4.5. MiRNA mature prediction methodology, c-mir-ch22.
a) MiRNA mature prediction methodology results. The first primer extension reaction results to a sharp band at 18nts which is the outcome of the extension of the precursor (denoted by P). The banding pattern on the second column is a consequence of the annealing of the second primer to the precursor or the mature (P/M) sequence. Also the banding pattern on the third column is the effect of the annealing of the third primer to the precursor or the mature sequence (P/M). b) Predicted potential mature miRNA sequences. Based on the results in (a) and the fact that the length of the expected mature is unknown [5] we predict two 21nts-long mature sequences. The black and red arrows indicate the predicted mature miRNA sequences based on the banding pattern of the 2nd and the 3rd primer extension reaction respectively. Northern blot analysis did not confirm any of these predictions.

To sum up, candidates c-mir-ch5a, c-mir-ch5b and c-mir-ch22, show several inconsistencies among the experiments. For c-mir-ch5a, the predicted mature based on the primer extension reaction was not verified via Northern blot analysis. On the contrary, a randomly selected sequence, which is in part

overlapping with the predicted mature, produced a band ~30 nts. In addition, the negative control probe, which was complement arm to the loop and a part of the opposite strand, 3', produced a band less than 25nts, Figure 4.3. Both results were unexpected in contrast with previous published data [5] and thus we were not able to resolve what is the mature sequence of that precursor.

According to previous experiments the expected mature sequence for c-mir-ch5b was 25nts long [5]. On the other hand, our experiments led to the conclusion that the mature length is 18 nts with 5'GCGAGGCUCGCGCGGUGCCCGCACC sequence, Figure 4.4.

For c-mir-ch22 we chose to set the length of the expected mature sequence at 21nts as it was unknown [5]. Along with the primer extension analysis we predict two possible matures, but none of them was confirmed via Northern blot analysis, Figure 4.5.

Finally, only for c-mir-ch9 our experiments were consistent with previous published data [5]. The mature sequence was 18nts long and was produced by the 5' strand of the precursor. We confirmed both by predicting its mature sequence via primer extension analysis and then we verified our prediction by Northern blot using DNA and LNA probes, Figure 4.2. For these reasons, the following experiments were performed only for c-mir-ch9.

Our goal was to investigate the expression levels of c-mir-ch9 in different cell lines. Northern blot analysis was performed on total RNA, which was kindly provided by Aristides Eliopoulos lab, IMBB, FORTH. The cell lines, which were evaluated, derived from different cancer types, bladder, lung, ovarian and melanoma. Several cell lines were examined for each cancer type, T24, VM cubi and EJ for bladder, SKME S1, H1299, HCC44 and H60 for lung, SKOV3 for ovarian and A375 for melanoma. In addition, HEK 293 cell line was also evaluated. As shown in Figure 4.6 and 4.7, c-mir-ch9 is expressed in bladder (T24, EJ) and lung (HCC44) cancer cell lines.

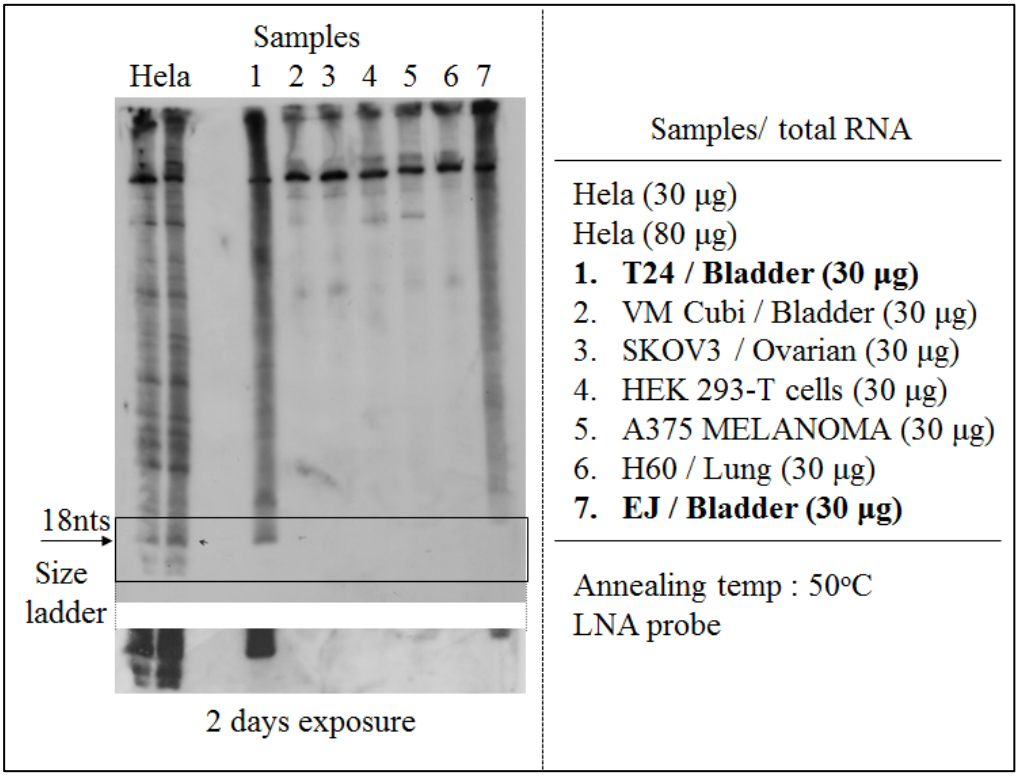

Figure 4.6. Northern blot analysis. The expression of c-mir-ch9 was evaluated on 30 µg of total RNA from different cell lines. T24, EJ and VM cubi derived from bladder cancer, SKOV3 ovarian cancer, HEK 293, A375 melanoma and H60 lung cancer. Hela cells were used in two quantities, 30µg and 80µg. In order to estimate the expression levels we employed the same LNA probe which was used during the evaluation of the mature sequence, Figure 4.2. As evident, only T24 and EJ express c-mir-ch9. Even after two days of exposure none of the other cell lines showed any level of expression.

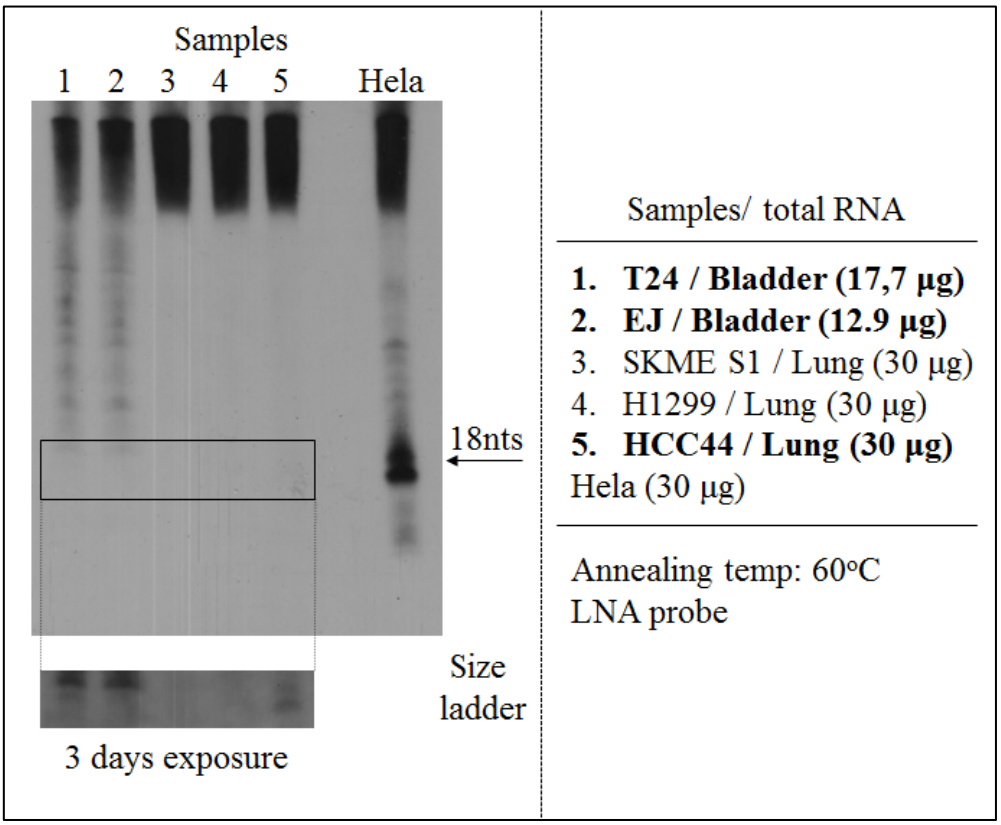

Figure 4.7. Northern blot analysis. The expression of c-mir-ch9 was evaluated on total RNA from different cell lines, T24, EJ derived from bladder cancer, SKME, H1299 and HCC44 lung cancer. The amount of the RNA used was 17.7µg and 12.9µg for T24 and EJ respectively and 30µg for the rest cell lines. In order to estimate the expression levels we utilized the same LNA probe which was used during the evaluation of the mature sequence, Figure 4.2. As evident from the figure, the expression of c-mir-ch9 is detectable in even more stringent conditions than in the previous experiment (Figure 4.6) since the annealing temperature was at 60°C. Moreover, HCC44 cell line express c-mir-ch9 in low levels as the corresponding band was obvious only after 3 days of exposure.

Following the experimental verification of the mature miRNA sequence of c-miR-ch9, we used TargetProfiler to scan all human 3'UTRs for potential targets of c-miR-ch9 [targetProfiler]. A total of 33 predicted targets for c-miR-ch9 achieved an HMM score of 6.2 (maximum score assigned by TargetProfiler =6.7) or higher (Table 4.1) and 17 of these where 8mers (as per Guo et al. [88]). One of these high scoring targets (HMM score: 6.2) was found to be located on a 3'UTR transcribed from chromosome 12. The miRNA::targetmRNA was an 8mer and displayed a low free energy (-23.70ΔG). Moreover, the seed was fully conserved in seven other organisms, excluding chimp. On selecting a miRNA target site for experimental verification it can be informative to obtain an intersection of predictions from other available target prediction tools. This target site was further confirmed by four other tools (TargetScan [99], StarMir [77], PITA [61], RNAhybrid [100]) which were used to perform target prediction using our novel miRNA sequence. The gene corresponding to this 3'UTR was CCND2, a gene with documented oncogenic activity [98] that is known to play a role in the G1/S transition of the cell cycle.

TABLE 4.1. c-miR-ch9 PREDICTED TARGETS.

Target Type	HMM Score	Location (chr:start-end)	Sequence (targetX&XmiRNA)	Bracket notation	Conservation score	Strand
7mer-m8	6.2	10:6019369-6019401	GAAGAGGACACCAGCCCAAGCUGGACCUGCCAUX&XCUGGCAGGGGGAGAGGUA	.(((....((...(((((((..&..)))))))...)).))).	7	1
8mer	6.2	11:26973747-26973779	GAGGUUCAAGGUGCUGCUUUGCAUGCCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....((((((...(.(((((((..&..)))))))..)..))))))	7	1
7mer-m8	6.2	1:156619755-156619787	AACCUGUCAGCUUGCACCAUCCCCACCUGCCACX&XCUGGCAGGGGGAGAGGUA	.....(((.((((((.(((((((..&..)))))))))).))))).	7	1
8mer	6.2	11:72794632-72794664	CUCAAAAGGUGAUUUUGUCCUUAGACCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(((.((((....(((((((..&..)))))))))).))))..	7	-1
7mer-m8	6.2	12:3771033-3771065	CCCGCUGUUAAACUGCAUAGGGCAGCCUGCCACX&XCUGGCAGGGGGAGAGGUA	.....(((((((..&..))))))).....	7	-1
8mer	6.2	12:4282963-4282995	AGUGGGGGCCGAGUUGUUCCCCCAGCCUGCCAAX&XCUGGCAGGGGGAGAGGUA	(((((..(((((((..&..)))))))..)).)).).))..	7	1
7mer-m8	6.2	1:35849529-35849561	GUAGGAGGUUUAGUGGCUGCUCUGGCCUGCCAUX&XCUGGCAGGGGGAGAGGUA	.....(((.((((..((((((..&..)))))))..)))).))).	7	-1
8mer	6.2	14:62243818-62243850	UAUUGCAUGUCCAGCUGGAUUCUGGCCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(((...((((..(((((((..&..)))))))))..))).	7	-1
7mer-m8	6.2	16:70239699-70239731	CAGCUGUUCUGUAUCAGUCCUACCACCUGCCAUX&XCUGGCAGGGGGAGAGGUA	.....(((.((.(((((((..&..)))))))...)))..	7	-1
8mer	6.2	17:44836546-44836578	UUGCAUCCUGCUGGGGCUGAACAUGCCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(((.....(.(((((((..&..)))))).)....))).	7	-1
8mer	6.2	19:44061675-44061707	AGCUUCCCCCAAGAAGUCCCCGCCACCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(((.((.(((((((..&..)))))))).).).)..	7	-1
8mer	6.2	19:45595285-45595317	AGGCAGCUGGUGGCUUUGCCCUCCACCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(((((((.(((((((..&..)))))))).)))).))))	7	-1
8mer	6.2	2:166362026-166362058	UGCCUACCUGUCAAACUGUGUGAAACCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(((.....(((((((..&..)))))))).....))).	7	1
7mer-m8	6.2	2:48517574-48517606	UGAAUUCGAGUAUUUUAAUGUUAUACCUGCCAUX&XCUGGCAGGGGGAGAGGUA	.....(((((((..&..))))))).....	7	1
7mer-m8	6.2	3:31653720-31653752	UAAAAGUGAAAGAGAAAGGGUUUUUCCUGCCACX&XCUGGCAGGGGGAGAGGUA	.....(((..(((((((..&..)))))))..))).....	7	1
8mer	6.2	4:114731538-114731570	CAGUAAAUAUAUUGAGCCAUGUUAACCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(((.(.((..(((((((..&..)))))))..)).)))).	7	-1
8mer	6.2	5:14561539-14561571	AGAAGUUCUUUCUCAUUCUCUUUCACCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(((((..(.(((((((..&..)))))))..))))))..	7	1
8mer	6.2	5:71536909-71536941	AUCUAGUUAAGUCGCUGAACAAUUACCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(((.....(((((((..&..)))))))......))).	7	1
8mer	6.2	8:92477834-92477866	GCAUCUAUAAAAGUAAAUUCUAGUGCCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....((((..(.(((((((..&..)))))))).).))))...	7	1
7mer-m8	6.2	X:40964086-40964118	UUCAUCUACUUAGACUUUUUAAAUGCCUGCCAUX&XCUGGCAGGGGGAGAGGUA	.....(((((((...(.(((((((..&..)))))))).))))))).	7	1
7mer-m8	6.2	X:44727602-44727634	AAUGCUGUUAUUUUUUCCAGAUUUACCUGCCAUX&XCUGGCAGGGGGAGAGGUA	.....((((((((((.....(((((((..&..))))))))))))))))	7	1

8mer	6.2	X:48787540-48787572	UUAUUGGGAGACUUUUGUCUUCCAGCCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(.((((((..(((((((..&..)))))))))))).))..	7	1
7mer-m8	6.7	1:157210240-157210272	GGCUGGGGAGUGUUUAUUUUAAGAUCCUGCCAUX&XCUGGCAGGGGGAGAGGUA	.....((((((....(((((((..&..)))))))...))))))	7	1
7mer-m8	6.7	1:157466940-157466972	UGGACUGUGCCUAUGGAUUUGGAUUCCUGCCAUX&XCUGGCAGGGGGAGAGGUA	.....(((((.....(((((((((..&..)))))))))...)))))	7	-1
8mer	6.7	1:208661990-208662022	AACAGUAACGAGUAGCCAGAGUACUCCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(((.....(((((((((..&..)))))))))....))).	7	1
8mer	6.7	16:28755701-28755733	CGUUCCCCAGGGGAGCUGGGGAAUUCCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(((.....(((((((((..&..)))))))))....))).	7	1
8mer	6.7	17:24980601-24980633	AAAUGGAGACUUCCAAUCACCAGCUCCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....(((..(((((((((..&..)))))))))....))).	7	-1
7mer-m8	6.7	19:55989973-55990005	GCCCGGCCUCCCGCCCAUGGGGUCUCCUGCCAUX&XCUGGCAGGGGGAGAGGUA	.....((((((((((..&..)))))))))....	7	1
7mer-m8	6.7	20:47955090-47955122	CCUGUUGGCUUGUGAAAUGAGCCCUCCUGCCACX&XCUGGCAGGGGGAGAGGUA	.....((((((((((..&..))))))))).....	7	-1
7mer-m8	6.7	21:15346087-15346119	GAAACCCAUUUAACUGUCACACACUCCUGCCACX&XCUGGCAGGGGGAGAGGUA	.....((.((.....((((((((((..&..)))))))))).))...	7	1
7mer-m8	6.7	5:172496015-172496047	AAAAAAAAUUGCAUUUUAUAUGAUUCCUGCCAUX&XCUGGCAGGGGGAGAGGUA	.....(((((((((..&..))))))))).....	7	1
8mer	6.7	5:65902001-65902033	UGAAUUUCUACGGAGCUUGAUGAUUCCUGCCAAX&XCUGGCAGGGGGAGAGGUA	.....((((.....(((((((((..&..)))))))...)))).	7	1
7mer-m8	6.7	8:74868493-74868525	GUAACAGGAAAAGUUUUCAUUAACUCCUGCCAUX&XCUGGCAGGGGGAGAGGUA	.....(((((((((..&..))))))))).....	7	-1

The table shows details for the total number of predicted targets for c-miRCh9. The target site on CCND2 is highlighted in red.

Experimental verification of the c-miR-ch9::CCND2 interaction

Since c-miR-ch9 was found in a genomic region that is frequently deleted in various cancer types and CCND2 has a documented oncogenic activity [98], the predicted interaction appears, at least in principle, quite plausible. Thus, we next performed experiments using reporter constructs carrying a firefly luciferase reporter to test whether the predicted interaction is functional. Given that many of the mammalian targets often contain binding sites (b.s.) for multiple miRNAs [76], we used constructs carrying binding sites that were repeated three times

(pGL4-10 + wt-Triplet) but also ~1,000 bp of the 3'UTR of CCND2 containing a single copy of the b.s. (pGL4-10 + wt-3'UTR). Moreover, constructs having mutations in the 5' seed site that disrupt the native pairing within the binding region of the triplet-cassette, as well as within the 3'UTR (designated as pGL4-10 + mut-Triplet and pGL4-10 + mut-3'UTR respectively) were also transfected, in order to provide a negative control. Furthermore, we performed transfection of empty vectors (pGL4-10) as a calibration control. All types of cassettes (constructs) prepared were placed into the pGL4-10 vector, downstream of the luc gene at XbaI site. HeLa cells were subsequently transfected with these reporter vectors carrying potential binding sites for c-miR-ch9. For every transfection assay, all constructs were tested in parallel: an empty luciferase vector (pGL4-10—Control), a wild-type triplet cassette containing potential binding sites for c-miR-ch9 (pGL4-10 + wt-Triplet), a wild-type 3'UTR containing a single copy of the potential b.s. for c-miR-ch9 (pGL4-10 + wt-3'UTR), a mutated triplet cassette containing binding sites with four point mutations (pGL4-10 + mut Triplet) and a mutated 3'UTR containing the same point mutations (pGL4-10 + mut-3'UTR). Since c-miR-ch9 was previously found to be expressed in HeLa cells at relatively high levels [5], there was no need for miRNA precursor overexpression. Firefly luciferase activity was measured and normalized against Renilla luciferase activity. The HeLa transfections were repeated 3 times using triplicate samples and the average relative expression is presented in Figure 5. The reporter construct carrying the wild-type CCND2 potential triplet binding sites (pGL4-10 + wt-Triplet) and the CCND2 wild-type-3'UTR (pGL4-10 + wt-3'UTR) appeared to be efficiently downregulated: the luciferase activity dropped to 49% (2.0 fold reduction-t-test: 1E-07) and 20% (5.0 fold reduction t-test:

2.22E-12), respectively, compared to 100% in the standardization control (pGL4-10—empty vector). We also assayed the mutated constructs pGL4-10 + mut-Triplet and pGL4-10 + mut-3'UTR, bearing CCND2 binding sites harbouring mutations in the “seed” element (at position 3, 4, 6 and 7—see Figure 4.1A). The transfection experiments confirmed that the downregulation previously observed was a result of the specific binding sites present in the wt-constructs. Luciferase expression was significantly increased both for the triplet mutated cassette (pGL4-10 + mut-Triplet) as well as the mutated-3'UTR (pGL4-10 + mut-3'UTR). This shows that miRNA-targeted regulation was suppressed due to truncated binding of the miRNA to the targets site(s). Specifically, in the case of pGL4-10 + mut-Triplet there was a ~2 fold increase (t-test: 2.09E-06) in luciferase activit y with respect to wt constructs (Figure 4.8A). Similarly for pGL4-10 + mut-3'UTR a 2.4 fold increase (t-test: 7.2E-06) was observed with respect to wt conditions (Figure 4.8B). It should be noted that, as expected, t-test analysis of pGL4-10 + mut-triplet expression vs. pGL4-10 expression showed that there was no significant difference between the expression of these two constructs (t-test: 0.663864). In the 3'UTR transfection assays, contrary to the triplet cassette assays, the levels of the pGL4-10 + mut-3'UTR expression did not achieve similar expression levels as in the pGL4-10 (empty) vector. One possible explanation for this is that by cloning a large portion of the CCND2 3'UTR we may have included other potential miRNA targets sites hence rendering this construct subject to additional regulation by other miRNAs. Three additional transfection experiments performed using the pGL4-10 + wt-Triplet constructs together with the anti-cmiR-Ch9 LNA inhibitor (25 nM) in order to block the predicted interaction of c-miR-ch9 with our reporter constructs, further

confirmed the true nature of this regulation. The co-transfection of anti-c-miR-ch9 LNA resulted in 1.5-fold increase (t-test: 0.004) in luciferase activity in the pGL4-10 + wt-Triplet-plusLNA transfected constructs with respect to the pGL4-10 + wt-Triplet (Figure 4.8C). Co-transfection of anti-c-miR-ch9 and pGL4-10 + wt-3'UTR was also performed and similar fold increase (1.5) in luciferase activity was observed in the pGL4-10 + wt-3'UTR-plus-LNA transfected constructs with respect to the pGL4-10 + wt-3'UTR (Figure 4.8E). An average for all the transfection experiments performed (total of 5 experiments with 3 triplicates for every condition) is shown in (Figure 4.8D). Although standard error bars show greater deviation from the mean in this summary of results, t-test analysis reveals that results remain statistically significant (pGL4-10 vs. pGL4-10 + wt-Triplet—t-test: 5.68E-10, pGL4-10 + wt-Triplet vs. pGL4-10 + mut-Triplet—ttest: 1.68E-05, pGL4-10 + wt-Triplet vs. pGL4-10 + wt-Tripletplus-LNA—t-test: 0.005805). As previously reported pGL4-10 vs. pGL4-10 + mut-triplet expression is not statistically significant (t-test: 0.221165). While confirmatory of the role of c-miRCh9 in targeting and regulating CCND2 targets sites, the lower luciferase expression observed for pGL4-10 + wt-Triplet-plusLNA with respect to pGL4-10 or pGL4-10 + mut-triplet also suggests the possible regulation of CCND2 by additional miRNAs in the same target site, which is in agreement with computational predictions (see Discussion).

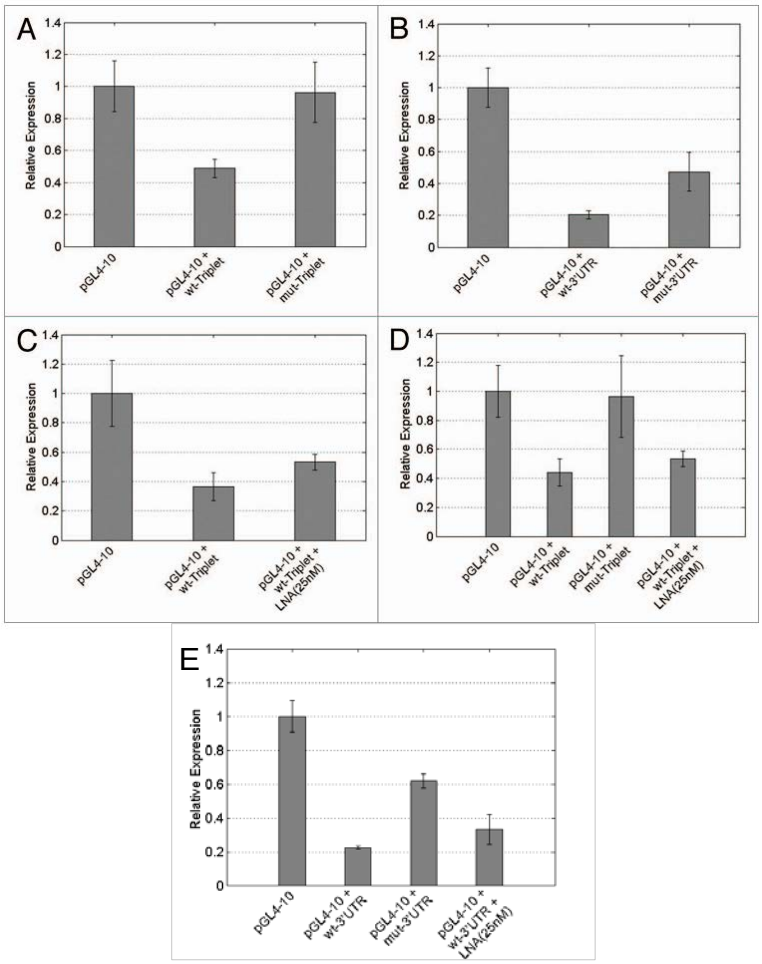

Figure 4.8. miRNA-sensor assay using luciferase expression as an indicator of miRNA activity after transfection of heLa cells with various constructs. (A) Relative luciferase expression after transfection of heLa cells with triplet-cassette constructs: pGL4-10—an empty pGL4-10 vector for standardization control, pGL4-10 + wt-Triplet—vector containing a wild-type triplet cassette containing potential binding sites for c-miR-ch9, pGL4-10 + mut-Triplet— a vector containing a triplet cassette with mutated binding sites for c-miR-ch9. (B) Relative luciferase expression after transfection of heLa cells with 3'UTR constructs. pGL4-10 — an empty pGL4-10 vector for standardization control, pGL4-10 + wt-3'UTR—vector containing a wild-type 3'UTR containing a single potential binding site for c-miR-ch9, pGL4-10 + mut-3'UTR—a vector containing a single mutated potential binding site for c-miR-ch9. (c) The pGL4-10 + wt-Triplet cassette transfection was repeated with concurrent addition of anti-LNA for our c-miR-ch9. (D) An average over all transfection experiments performed (total of 5 experiments with 3 triplicates for every condition). (E) Relative luciferase expression after transfection of HeLa cells with 3'UTR constructs: pGL4-10 – an empty pGL4-10 vector for standardization control, pGL4-10 + wt-3'UTR – vector containing a wild-type 3'UTR containing a single potential binding site for c-miR-ch9, pGL4-10 + mut-3'UTR – a vector containing a single mutated potential binding site for c-miR-ch9 and pGL4-10 + wt-3'UTR + LNA(25nM) - pGL4-10 + wt-3'UTR transfection was repeated with concurrent addition of anti-LNA for our c-miR-ch9 (3 triplicates where performed for every condition).

CONCLUDING REMARKS

We performed experimental verification on computational predictions of biological significance. In previous work [5] we showed the prediction and verification (via northern blot analysis) of 4 novel potential miRNA gene candidates.

As a follow-up to this work, using a modified version of primer extension analysis, we predicted the mature sequence of these miRNA genes. The mature sequences of, c-mir-ch9 and c-mir-ch5b, were also confirmed by Northern blot analysis. Due to the fact that several inconsistencies were observed among experiments as analyzed in chapter III, the rest of the experiments were performed only for c-miR-ch9.

First we evaluated the expression pattern of c-miR-ch9 in cell lines from different cancer types. The cell lines, which were examined, were T24, VM cubi and EJ for bladder cancer, SKME S1, H1299, HCC44 and H60 for lung, SKOV3 for ovarian and A375 for melanoma and HEK 293. C-miR-ch9 is expressed in bladder (T24, EJ) and lung (HCC44) cancer cell lines.

Thereafter, we utilized TargetProfiler to identify potential targets. The candidate under investigation, c-miR-ch9, is located in a cancer associated genomic region commonly deleted in various forms of bladder cancer [101]. Importantly, supporting evidence from recent deep sequencing studies do not report an expression for c-miR-Ch9 among the identified microRNA expression signatures of bladder cancer [102]. Computational identification of a highly significant and evolutionary conserved target binding site for this potential miRNA in the

CCND2 oncogene using Targetprofiler was the initial incentive for performing reporter gene assays.

CCND2 is a well-known cyclin which functions in the cell cycle and specifically in the G1/S transition. Moreover recent reporter assays have shown that CCND2 is targeted by let-7a and that this interaction inhibits proliferation in human prostate cancer cells both in vitro and in vivo[103]. Furthermore, bioinformatics analysis suggested that CCND2 is a putative target for miR-154. Subsequent experiments confirmed that miR-154 directly targets CCND2 in hepatocellular carcinoma (HCC), reduces tumorigenicity and inhibits the G1/S transition in cancer cells [98]. In line with these findings, our luciferase reporter assay results show that CCND2 is also targeted by c-miR-Ch9 as depicted by the decreased activity of the reporter gene in wild-type binding site conditions and the increased activity in mutated binding site conditions. Moreover, addition of anti-c-miR-Ch9 LNA to pGL4-10 + wt-Triplet conditions reduced regulation as shown by the observed increase in the reporter gene activity. However, luciferase activity in this case did not achieve the ~2-fold increase observed in the empty pGL4-10 vector or even the pGL4-10 + mut-Triplet constructs. One possible explanation for this is that other miRNA(s) compete for this target site. In fact, the target site under investigation is also a potential target site for 3 other known miRNAs (miR-182, miR-96 and miR-1271) as predicted by Targetprofiler as well as other target prediction tools (TargetScan, Diana-microT). However, using publicly available full genome tiling array [42] and next generation sequencing data [41] we observed that only miR-182 shows significant expression in HeLa cells. Competition between two miRNAs for the same target

site can explain our observed deviations in luciferase activity during LNA silencing of c-miR-Ch9.

CHAPTER V - DISCUSSION

CONCLUSIONS

This thesis has focused on two main areas:

I. The development of an algorithm, MiRduplexSVM, which is able to predict the mature molecule of a given miRNA pri-precursor.

II. The identification of the mature miRNA molecules of four newly identified miRNA genes and the verification of the functional interaction between one of these miRNAs and a cancer related target gene. The main innovations of this work are discussed in the following paragraphs.

MiRduplexSVM

We presented a novel methodology for the computational identification of the mature molecule(s) within novel miRNA hairpins. Our methodology takes into account several aspects of the biogenesis of miRNAs, whereby a duplex is formed before the mature molecule is selected. Our tool is the first that predicts miRNA duplexes and is shown to achieve much higher performance that four existing tools on both duplex and strand-specific miRNA prediction for mammalian hairpins. Moreover, the tool performs equally well on plant hairpins, without any particular customization.

MiRduplexSVM can be used to identify the miRNA:miRNA* duplex of a miRNA gene given the precursor sequence. The precursor sequence does not need to be precisely defined; it may be generated by one of the numerous computational tools that predict miRNA genes [5, 35, 38-40]. Such an approach is useful when searching for novel miRNAs that may be involved in a particular phenotype. For example, we recently developed and used a miRNA gene finding tool to locate potential new miRNAs residing is cancer associated genomic regions [5]. Similar efforts have been reported in a number of other studies, where new miRNA genes were computationally predicted [38, 104]. In order to verify that a predicted miRNA gene/precursor produces a functional miRNA however, a number of wet-lab experiments must be performed, requiring significant amount of time, money and effort [49]. MiRduplexSVM can provide reliable predictions about the most likely sequence of the miRNA molecule in these cases, thus guiding experimental efforts and ultimately reducing working hours and costs. Also MiRduplexSVM can be used to identify the mature molecule that lies on the opposite strand of a known miRNA performing better than formerly used methodologies [55, 57, 92, 93]. Another case where MiRduplexSVM would be useful is the in-silico study of factors that determine the cleavage sites of drosha and dicer, which define the miRNA:miRNA* duplex. This could be done by performing in silico mutagenesis experiments, generating predictions that can then guide the much more demanding wet - lab mutagenesis experiments [3, 6]. The final MiRduplexSVM model and the respective web server are available at http://139.91.171.154/duplexsvm/.

Regarding the experimental verification of mature miRNA molecules, we first identified the mature sequences of two possible miRNA genes, *c-mir-ch9*, *c-mir-ch5b*. We studied in more detail only *c-miR-Ch9* whose experimental results were consistent with previous published data[5]. We experimentally verified its expression in bladder and lung cancer's cell lines and its interaction with CCND2. The results reported here are important for two reasons: first, they confirm that our initial small RNA molecule shown by northern blot in [5] is indeed a true miRNA gene and second, that this miRNA targets and regulates CCND2.

Deletion of 09q33-34.1 (7MB) region, the region where the c-mir-ch9 is located, has been reported in several cancer types. In bladder cancer large tumors carried more frequently 9p deletions and tumors deleted in the regions 9ptr-p22, 9q22.3, 9q33, and 9q34 recurred significantly more rapidly than those without deletions [105]. In addition, loss of heterozygosity (LOH) on chromosome 9 is the most frequent genetic alteration identified in bladder tumors and is present in all stages and grades; approximately 60 ± 70% of bladder tumors show LOH of at least one locus on either arm of this chromosome [101]. Furthermore the 09q33-34.1 (7MB) region deletion has also been reported in lung[106] and ovarian cancer [107], but there was no association between tumour grade, stage or histopathology and any losses. Relating to bladder cancer, deletion of the 09q33-34.1 region refers to a LOH, implying that c-miR-Ch9 is expressed in lower levels than in wild type condition, which makes c-miR-Ch9 a probable tumor suppressor gene.

Furthermore, the role of CCND2 in proliferation, is another evidence that the recently discovered miRNA may function as a tumor suppressor[5]. Bladder cancer patients which exhibit deletion of the region retaining this miRNA may show increased proliferation by their inability to regulate CCND2, causing it to act like an oncogene and leading to failure of cells to arrest in G1/S and hence uncontrollable proliferation. In conclusion, our study used an integrative approach in which the prediction of a putative pre-miRNA is followed by the experimental verification of its mature sequence and the computational prediction of a target for this miRNA is experimentally confirmed using reporter assays. Our verified miRNA (c-mir-Ch9) was approved by the miRBase curation team and assigned the official miRNA name — hsa-mir-7150.

Overall, the most important contributions of this thesis are the development of MiRduplexSVM which is currently the most accurate and precise computational tool for the identification of the functional part of a miRNA gene (Chapter II) and the experimental verification of the mature sequence of has-mir-7150 and one of its targets showing that has-miR-7150 is a true miRNA and could probably act as a tumor suppressor molecule (Chapter IV). Smaller, yet important contributions include a comparative presentation of the strengths and limitations of existing mature miRNA prediction tools against a simple, naïve classifier which outperformed most of them (Chapter III).

FUTURE DIRECTIONS

MiRduplexSVM performs similarly to MaturePred in predicting mature miRNAs from plant precursors. Further optimization using plant precursors could lead to better results and is a direction I plan to explore in the future. In addition, the development of an interface, which combines miRNA gene, mature and target prediction algorithms, would be innovative in designing future biochemical studies which focus in the identification of uncharacterized miRNA genes. Future work will attempt to develop such a pipeline tool.

Another interesting future direction concerns the further study of hsa-miR-7150. An extensive investigation would shed new light on its cancer-related function and reveal the mechanisms of its operation. In vivo proliferation assays using bladder cancer cell lines and xenograph implantation using mice models are needed in order to explore this phenomenon.

REFERENCES

1. Calin, G.A., et al., *Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia.* Proc Natl Acad Sci U S A, 2002. **99**(24): p. 15524-9.
2. Cimmino, A., et al., *miR-15 and miR-16 induce apoptosis by targeting BCL2.* Proc Natl Acad Sci U S A, 2005. **102**(39): p. 13944-9.
3. Han, J., et al., *Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex.* Cell, 2006. **125**(5): p. 887-901.
4. Calin, G.A., et al., *Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers.* PNAS, 2004. **101**(9): p. 2999–3004.
5. Oulas, A., et al., *Prediction of novel microRNA genes in cancer-associated genomic regions--a combined computational and experimental approach.* Nucleic Acids Res, 2009. **37**(10): p. 3276-87.
6. Zeng, Y., R. Yi, and B.R. Cullen, *Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha.* EMBO J, 2005. **24**(1): p. 138-48.
7. Michael, M.Z., et al., *Reduced accumulation of specific microRNAs in colorectal neoplasia.* Mol Cancer Res, 2003. **1**(12): p. 882-91.
8. Kim, V.N., J. Han, and M.C. Siomi, *Biogenesis of small RNAs in animals.* Nat Rev Mol Cell Biol, 2009. **10**(2): p. 126-39.
9. Vermeulen, A., et al., *The contributions of dsRNA structure to Dicer specificity and efficiency.* Rna, 2005. **11**(5): p. 674-82.
10. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.
11. Carthew, R.W. and E.J. Sontheimer, *Origins and Mechanisms of miRNAs and siRNAs.* Cell, 2009. **136**(4): p. 642-55.
12. Gangaraju, V.K. and H. Lin, *MicroRNAs: key regulators of stem cells.* Nat Rev Mol Cell Biol, 2009. **10**(2): p. 116-25.
13. Nelson, P.T., W.X. Wang, and B.W. Rajeev, *MicroRNAs (miRNAs) in neurodegenerative diseases.* Brain Pathol, 2008. **18**(1): p. 130-8.
14. Thum, T., et al., *MicroRNA-21 contributes to myocardial disease by stimulating MAP kinase signalling in fibroblasts.* Nature, 2008. **456**(7224): p. 980-4.
15. Lu, J., et al., *MicroRNA expression profiles classify human cancers.* Nature, 2005. **435**(7043): p. 834-8.
16. Takamizawa, J., et al., *Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival.* Cancer Res, 2004. **64**(11): p. 3753-6.
17. Calin, G.A., et al., *Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia.* Proc Natl Acad Sci U S A, 2002 **99**(24): p. 15524–15529.
18. Hayashita, Y., et al., *A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation.* Cancer Res, 2005. **65**(21): p. 9628-32.

19. He, L., et al., *A microRNA polycistron as a potential human oncogene.* Nature, 2005. **435**(7043): p. 828-33.
20. Tagawa, H. and M. Seto, *A microRNA cluster as a target of genomic amplification in malignant lymphoma.* Leukemia, 2005. **19**(11): p. 2013-6.
21. Metzler, M., et al., *High Expression of precursor microRNA-155/BIC RNA in children with Burkitt lymphoma.* Genes Chromosomes Cancer, 2003. **39**(2): p. 167–169.
22. D'Arena, G., et al., *Biological and clinical heterogeneity of B-cell chronic lymphocytic leukemia.* Leuk Lymphoma, 2003. **44**(2): p. 223-8.
23. Baskerville, S. and D.P. Bartel., *Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes.* RNA . 2005(11): p. 241–247.
24. Jiang, Q., et al., *miR2Disease: a manually curated database for microRNA deregulation in human disease.* Nucleic Acids Res, 2009. **37**(Database issue): p. D98-104.
25. Lai, E.C., et al., *Computational identification of Drosophila microRNA genes.* Genome Biol, 2003. **4**(7): p. R42-61.
26. Lim, L.P., et al., *The microRNAs of Caenorhabditis elegans.* Genes Dev 2003b. **16**(8): p. 991–1008.
27. Weber, M.J., *New human and mouse microRNA genes found by homology search.* FEBS J., 2005. **272**(1): p. 59–73.
28. Legendre, M., A. Lambert, and D. Gautheret, *Profile-based detection of microRNA precursors in animal genomes.* Bioinformatics, 2004. **21**(7): p. 841–845.
29. Wang, X., et al., *MicroRNA identification based on sequence and structure alignment.* Bioinformatics, 2005. **21**(18): p. 3610–3614.
30. Buck, A.H., et al., *Discrete clusters of virus-encoded micrornas are associated with complementary strands of the genome and the 7.2-kilobase stable intron in murine cytomegalovirus.* J Virol, 2007. **81**(24): p. 13761-70.
31. Helvik, S.A., O. Snove, Jr., and P. Saetrom, *Reliable prediction of Drosha processing sites improves microRNA gene prediction.* Bioinformatics, 2006. **23**(2): p. 142-9.
32. Hertel, J. and P.F. Stadler, *Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data.* Bioinformatics, 2006. **22**(14): p. e197-202.
33. Sewer, A., et al., *Identification of clustered microRNAs using an ab initio prediction method.* BMC Bioinformatics, 2005. **6**: p. 267-281.
34. Xue, C., et al., *Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine.* BMC Bioinformatics, 2005. **6**: p. 310-316.
35. Batuwita, R. and V. Palade, *microPred: effective classification of pre-miRNAs for human miRNA gene prediction.* Bioinformatics, 2009. **25**(8): p. 989-95.
36. Ng, K.L. and S.K. Mishra, *De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures.* Bioinformatics, 2007. **23**(11): p. 1321-30.

37. Wang, M., et al., *New syntax to describe local continuous structure-sequence information for recognizing new pre-miRNAs.* J Theor Biol, 2010. **264**(2): p. 578-84.
38. Nam, J.W., et al., *Human microRNA prediction through a probabilistic co-learning model of sequence and structure.* Nucleic Acid Res, 2005. **33**(11): p. 3570-81.
39. Yousef, M., et al., *Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier.* Bioinformatics, 2006. **22**(11): p. 1325-34.
40. Terai, G., et al., *miRRim: a novel system to find conserved miRNAs with high sensitivity and specificity.* Rna, 2007. **13**(12): p. 2081-90.
41. Friedlander, M.R., et al., *Discovering microRNAs from deep sequencing data using miRDeep.* Nat Biotechnol, 2008. **26**(4): p. 407-15.
42. Kapranov, P., et al., *RNA maps reveal new RNA classes and a possible function for pervasive transcription.* Science, 2007. **316**(5830): p. 1484-8.
43. Landgraf, P., et al., *A mammalian microRNA expression atlas based on small RNA library sequencing.* Cell, 2007. **129**(7): p. 1401-14.
44. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.
45. Yousef, M., L. Showe, and M. Showe, *A study of microRNAs in silico and in vivo: bioinformatics approaches to microRNA discovery and target identification.* Febs J, 2009. **276**(8): p. 2150-6.
46. Pundhir, S. and J. Gorodkin, *MicroRNA discovery by similarity search to a database of RNA-seq profiles.* Front Genet, 2013. **4**: p. 133.
47. Gomes, C.P., et al., *A Review of Computational Tools in microRNA Discovery.* Front Genet, 2013. **4**: p. 81.
48. Allmer, J. and M. Yousef, *Computational methods for ab initio detection of microRNAs.* Front Genet, 2012. **3**: p. 209.
49. Oulas, A., et al., *A new microRNA target prediction tool identifies a novel interaction of a putative miRNA with CCND2.* RNA Biol, 2012. **9**(9): p. 1196-207.
50. Nam, J.W., et al., *Human microRNA prediction through a probabilistic co-learning model of sequence and structure.* Nucleic Acids Res, 2005. **33**(11): p. 3570-81.
51. Sheng, Y., P.G. Engstrom, and B. Lenhard, *Mammalian microRNA prediction through a support vector machine model of sequence and structure.* PLoS One, 2007. **2**(9): p. e946.
52. Helvik, S.A., O. Snove, Jr., and P. Saetrom, *Reliable prediction of Drosha processing sites improves microRNA gene prediction.* Bioinformatics, 2007. **23**(2): p. 142-9.
53. Nam, J.W., et al., *ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W455-8.
54. He, C., et al., *MiRmat: mature microRNA sequence prediction.* PLoS One, 2012. **7**(12): p. e51673.
55. Gkirtzou, K., et al., *MatureBayes: a probabilistic algorithm for identifying the mature miRNA within novel precursors.* PLoS One, 2010. **5**(8): p. e11843.

56. Wu, Y., et al., *MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences.* BMC Bioinformatics, 2011. **12**: p. 107.
57. Xuan, P., et al., *MaturePred: efficient identification of microRNAs within novel plant pre-miRNAs.* PLoS One, 2011. **6**(11): p. e27422.
58. Leclercq, M., A.B. Diallo, and M. Blanchette, *Computational prediction of the localization of microRNAs within their pre-miRNA.* Nucleic Acids Res, 2013.
59. Huttenhofer, A. and J. Vogel, *Experimental approaches to identify non-coding RNAs.* Nucleic Acids Res, 2006. **34**(2): p. 635-46.
60. Miranda, K.C., et al., *A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes.* Cell, 2006. **126**(6): p. 1203-17.
61. Kertesz, M., et al., *The role of site accessibility in microRNA target recognition.* Nat Genet, 2007. **39**(10): p. 1278-84.
62. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res, 2006. **34**(Database issue): p. D140-4.
63. Betel, D., et al., *The microRNA.org resource: targets and expression.* Nucleic Acids Res, 2008. **36**(Database issue): p. D149-53.
64. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.
65. Krek, A., et al., *Combinatorial microRNA target predictions.* Nat Genet, 2005. **37**(5): p. 495-500.
66. Maragkakis, M., et al., *DIANA-microT web server: elucidating microRNA functions through target prediction.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W273-6.
67. Maragkakis, M., et al., *Accurate microRNA target prediction correlates with protein repression levels.* BMC Bioinformatics, 2009. **10**: p. 295.
68. Khvorova, A., A. Reynolds, and S.D. Jayasena, *Functional siRNAs and miRNAs exhibit strand bias.* Cell 2003. **115**(2): p. 209–216.
69. Lee, Y., et al., *MicroRNA maturation: Stepwise processing and subcellular localization.* Embo Journal, 2002. **21**(17): p. 4663–4670.
70. Lim, L.P., et al., *Vertebrate microRNA genes.* Science, 2003. **299**(5612): p. 1540.
71. Brennecke, J., et al., *Principles of microRNA-target recognition.* PLoS Biol, 2005. **3**(3): p. e85.
72. Kiriakidou, M., et al., *A combined computational-experimental approach predicts human microRNA targets.* Genes Dev, 2004. **18**(10): p. 1165-78.
73. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs.* Genome Res, 2009. **19**(1): p. 92-105.
74. Witkos, T.M., E. Koscianska, and W.J. Krzyzosiak, *Practical Aspects of microRNA Target Prediction.* Curr Mol Med, 2011. **11**(2): p. 93-109.
75. Lewis, B.P., et al., *Prediction of mammalian microRNA targets.* Cell, 2003. **115**(7): p. 787-98.
76. Enright, A.J., et al., *MicroRNA targets in Drosophila.* Genome Biol, 2003. **5**(1): p. R1.
77. Long, D., et al., *Potent effect of target structure on microRNA function.* Nat Struct Mol Biol, 2007. **14**(4): p. 287-94.

78. Marin, R.M. and J. Vanicek, *Efficient use of accessibility in microRNA target prediction.* Nucleic Acids Res, 2011. **39**(1): p. 19-29.
79. Grimson, A., et al., *MicroRNA targeting specificity in mammals: determinants beyond seed pairing.* Mol Cell, 2007. **27**(1): p. 91-105.
80. Schmidt, T., H.W. Mewes, and V. Stumpflen, *A novel putative miRNA target enhancer signal.* PLoS One, 2009. **4**(7): p. e6473.
81. Baek, D., et al., *The impact of microRNAs on protein output.* Nature, 2008. **455**(7209): p. 64-71.
82. Selbach, M., et al., *Widespread changes in protein synthesis induced by microRNAs.* Nature, 2008. **455**(7209): p. 58-63.
83. Mayr, C., M.T. Hemann, and D.P. Bartel, *Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation.* Science, 2007. **315**(5818): p. 1576-9.
84. Sylvestre, Y., et al., *An E2F/miR-20a autoregulatory feedback loop.* J Biol Chem, 2007. **282**(4): p. 2135-43.
85. Papadopoulos, G.L., et al., *The database of experimentally supported targets: a functional update of TarBase.* Nucleic Acids Res, 2009. **37**(Database issue): p. D155-8.
86. Lee, Y., et al., *Network modeling identifies molecular functions targeted by miR-204 to suppress head and neck tumor metastasis.* PLoS Comput Biol, 2010. **6**(4): p. e1000730.
87. Lorenz, R., et al., *ViennaRNA Package 2.0.* Algorithms Mol Biol, 2011. **6**: p. 26.
88. Karathanasis, N. *SVM-based miRNA: MiRNA∗ duplex prediction*. 2012.
89. Hamelryck, T., *Probabilistic models and machine learning in structural bioinformatics.* Stat Methods Med Res, 2009. **18**(5): p. 505-26.
90. Kozomara, A. and S. Griffiths-Jones, *miRBase: integrating microRNA annotation and deep-sequencing data.* Nucleic Acids Res. **39**(Database issue): p. D152-7.
91. Chang, C. and C. Lin, *LIBSVM: A library for support vector machines.* ACM Trans. Intell. Syst. Technol., 2011. **2**(3).
92. Weber, M.J., *New human and mouse microRNA genes found by homology search.* Febs J, 2005. **272**(1): p. 59-73.
93. Artzi, S., A. Kiezun, and N. Shomron, *miRNAminer: a tool for homologous microRNA gene search.* BMC Bioinformatics, 2008. **9**: p. 39.
94. Krol, J., et al., *Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design.* J Biol Chem, 2004. **279**(40): p. 42230-9.
95. Zhang, X. and Y. Zeng, *The terminal loop region controls microRNA processing by Drosha and Dicer.* Nucleic Acids Res, 2010. **38**(21): p. 7689-97.
96. Auyeung, V.C., et al., *Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing.* Cell, 2013. **152**(4): p. 844-58.
97. Larranaga, P., et al., *Machine learning in bioinformatics.* Brief Bioinform, 2006. **7**(1): p. 86-112.

98. Wang, W., et al., *Human tumor microRNA signatures derived from large-scale oligonucleotide microarray datasets.* Int J Cancer, 2011. **129**(7): p. 1624-34.
99. Lewis, B.P., Burge, C.B., and Bartel, D.P. , *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell 2005(120): p. 15-20.
100. Rehmsmeier, M., et al., *Fast and effective prediction of microRNA/target duplexes.* Rna, 2004. **10**(10): p. 1507-17.
101. Simoneau, M., et al., *Four tumor suppressor loci on chromosome 9q in bladder cancer: evidence for two novel candidate regions at 9q22.3 and 9q31.* Oncogene, 1999. **18**(1): p. 157-63.
102. Han, Y., et al., *MicroRNA expression signatures of bladder cancer revealed by deep sequencing.* PLoS One, 2011. **6**(3): p. e18286.
103. Dong, Q., et al., *MicroRNA let-7a inhibits proliferation of human prostate cancer cells in vitro and in vivo by targeting E2F2 and CCND2.* PLoS One. **5**(4): p. e10147.
104. Bentwich, I., et al., *Identification of hundreds of conserved and nonconserved human microRNAs.* Nat Genet, 2005. **37**(7): p. 766-70.
105. Simoneau, M., et al., *Chromosome 9 deletions and recurrence of superficial bladder cancer: identification of four regions of prognostic interest.* Oncogene, 2000. **19**(54): p. 6317-23.
106. Fong, Y., et al., *Chromosomal imbalances in lung adenocarcinomas with or without mutations in the epidermal growth factor receptor gene.* Respirology, 2010. **15**(4): p. 700-5.
107. Devlin, J., et al., *High frequency of chromosome 9 deletion in ovarian cancer: evidence for three tumour-suppressor loci.* Br J Cancer, 1996. **73**(4): p. 420-3.

APPENDIX

MISSING DUPLEXES PREDICTIONS OF HUMAN AND MOUSE HAIRPINS

>hsa-mir-95_5p
CUCAAUAAAUGUCUGUUGAAU
>hsa-mir-103a-1_5p
GGCUUCUUUACAGUGCUGCCUUG
>hsa-mir-107_5p
AGCUUCUUUACAGUGUUGCCUUG
>hsa-mir-208a_5p
GAGCUUUUGGCCCGGGUUAUAC
>hsa-mir-147a_5p
GACAACAUUUCUGCACACACAC
>hsa-mir-203a_5p
AGUGGUUCUUAACAGUUCAACA
>hsa-mir-210_5p
AGCCCCUGCCCACCGCACACUG
>hsa-mir-1-2_5p
ACAUACUUCUUUAUGUACCCAU
>hsa-mir-128-1_5p
CGGGGCCGUAGCACUGUCUGA
>hsa-mir-133a-1_5p
AGCUGGUAAAAUGGAACCAAAU
>hsa-mir-133a-2_5p
AGCUGGUAAAAUGGAACCAAAU
>hsa-mir-137_5p
ACGGGUAUUCUUGGGUGGAUAAU
>hsa-mir-152_5p
AGGUUCUGUGAUACACUCCGACU
>hsa-mir-153-1_5p
GUCAUUUUUGUGAUCUGCAGCU
>hsa-mir-153-2_5p
UCAUUUUUGUGAUGUUGCAGCU
>hsa-mir-184_5p
CCUUAUCACUUUUCCAGCCCAGC
>hsa-mir-206_5p
ACAUGCUUCUUUAUAUCCCCAU
>hsa-mir-320a_5p
GCCUUCUCUUCCCGGUUCUUCC
>hsa-mir-1-1_5p
ACAUACUUCUUUAUAUGCCCAU
>hsa-mir-128-2_5p
GGGGGCCGAUACACUGUACGAGA
>hsa-mir-101-2_5p
UCGGUUAUCAUGGUACCGAUGCU
>hsa-mir-370_5p
CAGGUCACGUCUCUGCAGUUAC
>hsa-mir-372_5p
CCUCAAAUGUGGAGCACUAUUC
>hsa-mir-375_5p
GCGACGAGCCCCUCGCACAAAC
>hsa-mir-328_5p
GGGGGGCAGGAGGGGCUCAGGG
>hsa-mir-326_5p
GGAGGCAGGGCCUUUGUGAAGGCG
>hsa-mir-133b_5p
GCUGGUCAAACGGAACCAAGUC
>hsa-mir-384_5p
UAAACAAUUCCUAGACAAUAUG
>hsa-mir-448_5p
GAACAUCCUGCAUAGUGCUGCC
>hsa-mir-429_5p
CGUCUUACCAGACAUGGUUAGA
>hsa-mir-433_5p
UACGGUGAGCCUGUCAUUAUUC
>hsa-mir-329-1_5p
AGAGGUUUUCUGGGUUUCUGUUU
>hsa-mir-329-2_5p
AGAGGUUUUCUGGGUUUCUGUUU
>hsa-mir-412_5p
GAUGGUCGACCAGUUGGAAAGU
>hsa-mir-410_5p
AGGUUGUCUGUGAUGAGUUCGC
>hsa-mir-487a_5p
AGUGGUUAUCCCUGCUGUGUUC
>hsa-mir-489_5p
UGGUCGUAUGUGUGACGCCAUU
>hsa-mir-494_5p
AGGUUGUCCGUGUUGUCUUCUC
>hsa-mir-496_5p
AGGUUGUCCAUGGUGUGUUCA
>hsa-mir-520e_5p
CUCAAGAUGGAAGCAGUUUCUG
>hsa-mir-520f_5p
CUCUAAAGGGAAGCGCUUUCUG
>hsa-mir-520b_5p
CCUCUACAGGGAAGCGCUUUCU
>hsa-mir-518b_5p
CUCCAGAGGGAAGCGCUUUCUG
>hsa-mir-519d_5p
CUCCAAAGGGAAGCGCUUUCUG
>hsa-mir-521-2_5p
CUCCAAAGGGAAGAAUUUUCUC
>hsa-mir-520g_5p
ACCCUCUAGAGGAAGCACUUU
>hsa-mir-520h_5p
UAGAGGAAGCACUUUCUGUUUG
>hsa-mir-521-1_5p
CCUCCAAAGGGAAGAACUUUCU
>hsa-mir-519a-2_5p
CUCUACAGGGAAGCGCUUUCUG
>hsa-mir-507_5p

CACUUCAAGAAGUGCCAUGCAU
>hsa-mir-544a_5p
UCUUGUUAAAAAGCAGAUUCU
>hsa-mir-487b_5p
UGGUUAUCCCUGUCCUGUUCG
>hsa-mir-551a_5p
GAAAUCCAGAGUGGGUGGGGCC
>hsa-mir-552_5p
UGUUUAACCUUUUGCCUGUUGG
>hsa-mir-555_5p
UCAGAUGUGGAGCACUACCUUU
>hsa-mir-557_5p
UGGAGGCCUGGGGCCCUCCCUG
>hsa-mir-558_5p
UUUUGGUAUAGUAGCUCUAGA
>hsa-mir-562_5p
AUAUGGUCAGUCUACUUUUAG
>hsa-mir-563_5p
UAGGAAAUGUGUGUUGCUCUG
>hsa-mir-569_5p
UUUGUGGGACAUUAACAACAGC
>hsa-mir-571_5p
CUCAGUGUGCCAUUUCCUUGUC
>hsa-mir-572_5p
GGGGCCGCUGCGGGCGGAAGG
>hsa-mir-575_5p
AAUUCAGCCCUGCCACUGGCUU
>hsa-mir-578_5p
CAAUCCCGGACAACAAGAAGCU
>hsa-mir-579_5p
UCGCGGUUUGUGCCAGAUGAC
>hsa-mir-580_5p
UAAUGAUUCAUCAGACUCAGAU
>hsa-mir-585_5p
CCUAGCACACAGAUACGCCCAG
>hsa-mir-548a-1_5p
AAAAGUAAUUGUGAUUUUUGCC
>hsa-mir-548a-2_5p
AAAAGUAAUUGGGGUUUUUUGC
>hsa-mir-595_5p
UUAACACCAGCACGCUCAAUGUAG
>hsa-mir-598_5p
GCGGUGAUCCCGAUGGUGUGAGC
>hsa-mir-599_5p
UUGAUAAGCUGACAUGGGACAG
>hsa-mir-600_5p
GGAAGGCUCUUGUCUGUCAGG
>hsa-mir-603_5p
AAAAGUAAUUGCAGUGCUUCCC
>hsa-mir-604_5p
UGCUUGACCUUCCACGCUCUCG
>hsa-mir-606_5p
CCUUGGUUUUUAGUAGUUUUAC
>hsa-mir-607_5p
UAUAGAUCUGGAUUGGAACCCA
>hsa-mir-611_5p
UGAGAGCGUUGAGGGGAGUUCCAG
>hsa-mir-613_5p
UGAAGGGACCCUUCCUGUAGU
>hsa-mir-614_5p
UCUGAAGCCUGCAGGGGCAGGC
>hsa-mir-619_5p
UACAGGCAUGAGCCACUGCGG
>hsa-mir-620_5p
AUCUAUAUCUAGCUCCGUAUAU
>hsa-mir-621_5p
GGUAGGCGGUGCUGCUGUGCUC
>hsa-mir-622_5p
UACUGGUCUCAGCAGAUUGAGG
>hsa-mir-626_5p
GGAGUAUUUUUAUGCAAUCUGA
>hsa-mir-630_5p
CCUCUUUGUAUCAUAUUUUGU
>hsa-mir-632_5p
UGACGGGAGGCGGAGCGGGGAA
>hsa-mir-633_5p
UUGCGGUAGAUACUAUUAACC
>hsa-mir-634_5p
AUCGAGGGUUGGGGCUUGGUGU
>hsa-mir-636_5p
CGCGGGCGGGGCCGGCCCCGCU
>hsa-mir-637_5p
UCGGGCUCCCCACUGCAGUUAC
>hsa-mir-639_5p
GGGGCGCGCGCGGCCUGGAGGG
>hsa-mir-640_5p
UUCCUGAAGAUCAGACACAUC
>hsa-mir-643_5p
ACCUGAGCUAGAAUACAAGUAG
>hsa-mir-644a_5p
UCAUAAGGAAUGUUGCUCUG
>hsa-mir-645_5p
AGACCAGUACCGGUCUGUGGCCU
>hsa-mir-646_5p
GGAGUCAGCACACCUGCUUUUC
>hsa-mir-649_5p
UUUUUGAUCGACAUUUGGUUGAA
>hsa-mir-661_5p
GGGGCAGGCGCAGGCCUGAGCCC
>hsa-mir-662_5p
GCCAGGCCCUGACGGUGGGGUGG
>hsa-mir-655_5p
AGAGGUUAUCCGUGUUAUGUUC
>hsa-mir-656_5p
AGGUUGCCUGUGAGGUGUUCACU
>hsa-mir-549a_5p
AGCUCAUCCAUAGUUGUCACU
>hsa-mir-658_5p
GGGCCCUGCCCGCCCGCCAGCU
>hsa-mir-421_5p
CCUCAUUAAAUGUUUGUUGAAUGA
>hsa-mir-1264_5p
AGGUCCUCAAUAAGUAUUUGUU
>hsa-mir-668_5p
GUAAGUGCGCCUCGGGUGAGCAUG
>hsa-mir-151b_5p
UCUCUUCAGGGCUCCCGAGACA
>hsa-mir-320b-1_5p

UCCCUCUCUUUCUAGUUCUUCC
>hsa-mir-320c-1_5p
GCCUUCUCUUCCCAGUUCUUCC
>hsa-mir-1301_5p
GGUCGCUCUAGGCACCGCAGCA
>hsa-mir-320b-2_5p
AGGCUUUCUCUUCCCAGAUUUCC
>hsa-mir-762_5p
UCUCGGCCCGUACAGUCCGGC
>hsa-mir-2113_5p
CAAUGUGUGACAGGUACAGGGA
>hsa-mir-765_5p
AGUAGACAGCCCUUUUCAAGCC
>hsa-mir-300_5p
AGAGAGGUAAUCCUUCACGCAU
>hsa-mir-892a_5p
UACUCAGAAAGGUGCCAGUCAC
>hsa-mir-874_5p
CGGCCCCACGCACCAGGGUAAG
>hsa-mir-892b_5p
UACUCAGAAAGGUGCCAUUUAU
>hsa-mir-889_5p
AAUGGCUGUCCGUAGUAUGGUC
>hsa-mir-147b_5p
UGGAAACAUUUCUGCACAAACUAG
>hsa-mir-887_5p
CUUGGGAGCCCUGUUAGACUC
>hsa-mir-665_5p
AGGGGUCUCUGCCUCUACCCAG
>hsa-mir-543_5p
AAGUUGCCCGUGUUUUUUUCG
>hsa-mir-760_5p
CCCCUCAGUCCACCAGAGCCCGG
>hsa-mir-301b_5p
GCUCUGACGAGGUUGCACUACU
>hsa-mir-208b_5p
AAGCUUUUUGCUCGAAUUAUGU
>hsa-mir-920_5p
UAGUUGUUCUACAGAAGACC
>hsa-mir-922_5p
UCCCUCUCCCUGUCCUGGACUG
>hsa-mir-933_5p
AGAGGUCCUCGGGGCGCGCGUC
>hsa-mir-935_5p
AGUGGCGGGAGCGGCCCCUCG
>hsa-mir-940_5p
AGGAGCGGGGCCUGGGCAGCCC
>hsa-mir-941-1_5p
ACAUGUGCCCAGGGCCCGGGACA
>hsa-mir-941-2_5p
ACAUGUGCCCAGGGCCCGGGACA
>hsa-mir-941-3_5p
ACAUGUGCCCAGGGCCCGGGACA
>hsa-mir-941-4_5p
ACAUGUGCCCAGGGCCCGGGACA
>hsa-mir-943_5p
UGGGGGACGUUUGCCGGUCACU
>hsa-mir-944_5p
CAUCUGAUAUACAAUAUUUUCU
>hsa-mir-1180_5p
UGGACCCACCCGGCCGGGAAUA
>hsa-mir-1182_5p
UCUCCUCCCCUCUCCAGCAGCGA
>hsa-mir-1183_5p
ACACAGAACAUUAGAGAAGAC
>hsa-mir-1184-1_5p
UUCUGCUCAGCAGUCAACAGUG
>hsa-mir-663b_5p
CGAGGGCCGUCCGGCAUCCUAG
>hsa-mir-548e_5p
CAAAAGCAAUCGCGGUUUUUGC
>hsa-mir-1285-2_5p
UGCAUCACUUGAGCCCAGCAAU
>hsa-mir-1286_5p
GGGACUCAGCUUGCUCUGGCU
>hsa-mir-1289-1_5p
AUGCAGACUCUUGGUUUCCACCCC
>hsa-mir-1289-2_5p
AAGGCACAUUCCUAGACCCUGC
>hsa-mir-1290_5p
GAGCGUCACGUUGACACUCAAA
>hsa-mir-1295a_5p
GCCCAGAUCCGUGGCCUAUUC
>hsa-mir-1297_5p
UAGGGUUGAUCUAUUAGAAUUA
>hsa-mir-1299_5p
CCUCAUGGCAGUGUUCUGGAAUCC
>hsa-mir-1302-1_5p
UGUAUGUAAGAAUAUCCCAUAC
>hsa-mir-1302-2_5p
UAGCAUAAAUAUGUCCCAAGC
>hsa-mir-1302-3_5p
UAGCAUAAAUAUUUCCCAAGC
>hsa-mir-1302-4_5p
UUAGAAUAAGUAUGUCUCCAUG
>hsa-mir-1302-5_5p
UAGGUAUAAGUAUAUCCCAUGU
>hsa-mir-1302-6_5p
UUGGUAAUAUAUGUAUGGCCCAC
>hsa-mir-1302-7_5p
UAGGACAUGUAUGUCUGGUGC
>hsa-mir-1302-8_5p
UUUUCAGCAUUAGUGUAUCACA
>hsa-mir-1303_5p
AGCGAGACCUCAACUCUACAAU
>hsa-mir-1305_5p
UCUACCAUUAGUUUUGAAUGUU
>hsa-mir-548f-1_5p
UGCAAAAGUAAUCACAGUUUUU
>hsa-mir-548f-2_5p
CGAACAUAAUUGCAGUUUUUAU
>hsa-mir-548f-3_5p
AAACCUAAUUGCAAUUUUUGC
>hsa-mir-548f-4_5p
AAAAGUAAUAGUGGUUUUUGCC
>hsa-mir-548f-5_5p
AAAGUAAUCAUGUUUUUUUCC
>hsa-mir-1244-1_5p

AUCUUAUUCCGAGCAUUCCAG
>hsa-mir-1245a_5p
UAUAGGCCUUUAGAUCAUCUGA
>hsa-mir-1249_5p
AGGAGGGAGGAGAUGGGCCAAGUUC
>hsa-mir-1258_5p
UCCACGACCUAAUCCUAACUCC
>hsa-mir-548o_5p
AAAAUGUGUUGAUUGUAAUGGU
>hsa-mir-1269a_5p
AGUUGGCAUGGCUCAGUCCAAGU
>hsa-mir-1273a_5p
UGAGGCAGGAGAAUUGCUUGA
>hsa-mir-302f_5p
CUGUGUAAACCUGGCAAUUUUC
>hsa-mir-548p_5p
UUAAUUGCAGUUUUUGUCAUU
>hsa-mir-1278_5p
AGAUGAUAUGCAUAGUACUCCC
>hsa-mir-1281_5p
AGGGGGCACCGGGAGGAGGUG
>hsa-mir-1288_5p
AGCAGAUCAGGACUGUAACUC
>hsa-mir-1321_5p
CAAGUAUUAUUAUCCCUGUUUU
>hsa-mir-1322_5p
AGUAUCAUGAAUUAGAAACCU
>hsa-mir-1197_5p
CGGUUGACCAUGGUGUGUACG
>hsa-mir-1324_5p
UGCAUGAAGCCUGGUCCUGCCC
>hsa-mir-1537_5p
AGCUGUAAUUAGUCAGUUUUCU
>hsa-mir-1538_5p
ACAGCAGCAACAUGGGCCUCG
>hsa-mir-1539_5p
GGCUCUGCGGCCUGCAGGUAG
>hsa-mir-320d-1_5p
UUCUCGUCCCAGUUCUUCC
>hsa-mir-320c-2_5p
CUUCUCUUUCCAGUUCUUCC
>hsa-mir-320d-2_5p
UUCUCUUCCCAGUUCUUC
>hsa-mir-1825_5p
AGAGACUGGGGUGCUGGGCU
>hsa-mir-1827_5p
UCAGCAGCACAGCCUUCAG
>hsa-mir-1912_5p
UGCUCAUUGCAUGGGCUGUGUA
>hsa-mir-1913_5p
CGGCAGAGGAGGCUGCAGAGGC
>hsa-mir-1972-1_5p
UGCCACCACACCUGGCUUAAAU
>hsa-mir-1973_5p
UAUGUUCAACGGCCAUGGUAU
>hsa-mir-1976_5p
CAGCAAGGAAGGCAGGGGUC
>hsa-mir-2053_5p
AGAUUUAAUUAACAUUUGCAACC
>hsa-mir-2117_5p
UCUGUCCGGCAUGGUGAACAGC
>hsa-mir-2276_5p
GCCCUCUGUCACCUUGCAGACG
>hsa-mir-711_5p
AGUCUCUCCUCAGGGUGCUGC
>hsa-mir-718_5p
GCGCGCAAGAUGGCGGCGGGCC
>hsa-mir-2861_5p
UCCGGCUCCCCCUGGCCUCUC
>hsa-mir-3116-1_5p
UCCCUACUAUGUUCCAGGCACU
>hsa-mir-3116-2_5p
UCCCUACUAUGUUCCAGGCACCU
>hsa-mir-3118-1_5p
AAUUUUCAUAAUGCAAUCACAC
>hsa-mir-3118-2_5p
AAUUUUCAUAAUGCAAUCACAC
>hsa-mir-3118-3_5p
AAUUUUCAUAAUGCAAUCACAC
>hsa-mir-3123_5p
UUGAAUGAUUCUCCCAUUUC
>hsa-mir-548s_5p
UAAUUGCAGUUUUUGCCAUUAU
>hsa-mir-378b_5p
AUUGAGUCUUCAAGGCUAGUG
>hsa-mir-3134_5p
UGUGUAGUCUUUUAUCCCUCACA
>hsa-mir-466_5p
UGUGUUGCAUGUGUGUAUAUGU
>hsa-mir-544b_5p
UAAAAUGCAGAAUCCAUUUCU
>hsa-mir-3138_5p
ACUUCCCCCACCUCACUGCCCG
>hsa-mir-3142_5p
UGAACCUUCAGAAAGGCUGCUG
>hsa-mir-548u_5p
AAAAGUAAUGUGGUUUUUUUC
>hsa-mir-3146_5p
CUUCUUUCUAUCCUAGUAUAAC
>hsa-mir-548v_5p
UGAGCAAAAGUAAUUGCGGUUUU
>hsa-mir-3149_5p
ACACACACAUGUCAUCCACACA
>hsa-mir-3153_5p
AAUGUCCCUGUCCCCUUCCCCC
>hsa-mir-3154_5p
AGCCCCAGCUCCCGCUCACCC
>hsa-mir-3155a_5p
CUCCCACUGCAGAGCCUGGGG
>hsa-mir-3166_5p
AGGCAUUGUCUGCGUUAGGAUU
>hsa-mir-3167_5p
ACCAGUAUUUCUGAAAUUCUU
>hsa-mir-3118-4_5p
AAUUUUCAUAAUGCAAUCACAC
>hsa-mir-3176_5p
CGGCAGCCUCGGGCCACACUCC
>hsa-mir-3179-1_5p

GUUUAAAUUACACUCCUUCUGC
>hsa-mir-3179-2_5p
GUUUAAAUUACACUCCUUCUGC
>hsa-mir-3179-3_5p
GUUUAAAUUACACUCCUUCUGC
>hsa-mir-3188_5p
CUGCUGUGCCGCCAGGGCCUCC
>hsa-mir-320e_5p
GCCUUCUCUUCCCAGUUCUUC
>hsa-mir-3118-5_5p
AAUAAUAUUCAUAAUGCAAUCA
>hsa-mir-3198-1_5p
UCACUGUUCACCCAGCACUAG
>hsa-mir-4293_5p
GUUCCUUGGGAAGCUGGUGACA
>hsa-mir-4299_5p
UGACCAAUCAUGUUACAGUGUU
>hsa-mir-4300_5p
AGAGGGCCAGCUAAAUCAGCAG
>hsa-mir-4306_5p
UAGUGUCCUUAGAGUCUCCAGA
>hsa-mir-4307_5p
UCAGAAGAAAAAAACAGGAGAU
>hsa-mir-4308_5p
UCAGAGGGAACUCCAUUGGAC
>hsa-mir-4310_5p
CGUCUGGGGCCUGAGGCUGCAG
>hsa-mir-4312_5p
GGGCACAGAGAGCAAGGAGCC
>hsa-mir-4318_5p
UUAUGUCAUAAACCCACUGUG
>hsa-mir-4320_5p
UGGGGUUUGCUGUAGACAUUUC
>hsa-mir-4322_5p
AGUUCCGCGCCUGGCCGUGU
>hsa-mir-4321_5p
AGAGCCUCUGCCCCUCCCGAGA
>hsa-mir-4323_5p
CAGGCGGGCAUGUGGGGUGUC
>hsa-mir-4324_5p
UAAGGGUCUCAGCUCCAGGGAA
>hsa-mir-4257_5p
CAGUCCCUAGGUAGGAUUUGGGG
>hsa-mir-4259_5p
UGUGUCCUGAAUUGGGUGGGG
>hsa-mir-4253_5p
AUCGCCCUUGAGGGGCCCU
>hsa-mir-4251_5p
CGUCCUCCAGCUUUUUUCCUUA
>hsa-mir-4254_5p
AGGGUGGGGUGGCUCCUCUGCA
>hsa-mir-4252_5p
CUGGCAGCUCAUCAGUCCAG
>hsa-mir-4261_5p
UGGAAGUGGGUUCCUCCCAGU
>hsa-mir-4265_5p
UGGAGCUUCAGCCUACACCU
>hsa-mir-4266_5p
CUGCUGGCCGGGGCCCCUACUC
>hsa-mir-4262_5p
AAGCUGCAGGUGCUGAUGUUGG
>hsa-mir-4268_5p
ACAUCAGGUUCUAGAGGUUUU
>hsa-mir-4263_5p
UCAGGGUUUUACUUGGGAGAUUGG
>hsa-mir-4271_5p
CUCCAUAUCUUUCCUGCAGCC
>hsa-mir-4272_5p
UGCACAAAUUAAUCAGUUAAU
>hsa-mir-4274_5p
AGGGUAACUGAGCUGCUGCCGG
>hsa-mir-4281_5p
CCCCCGACAGUGUGGAGCUGGG
>hsa-mir-4279_5p
CUCUGUGGAGCUGAGGAGCA
>hsa-mir-4278_5p
AGGAGAAUCCCAUAGAACAUU
>hsa-mir-4282_5p
AAGUUCCAGGGGAAGAUUUUAGU
>hsa-mir-4288_5p
AGAGUCAUCAGCAGCACU
>hsa-mir-4292_5p
CCUGCUUAGGAGGCCAGAGGUG
>hsa-mir-4290_5p
AAGGUGAAGGGAGGGUCAGU
>hsa-mir-4329_5p
AGGUGUACCAGGGUUUUGGAGU
>hsa-mir-4330_5p
AGGCAAUUAUCUGAGGAUGCAG
>hsa-mir-4328_5p
CAGUUGAGUCCUGAGAACCAUUG
>hsa-mir-1184-2_5p
UUCUGCUCAGCAGUCAACAGUG
>hsa-mir-1184-3_5p
UUCUGCUCAGCAGUCAACAGUG
>hsa-mir-1233-2_5p
AGUGGGAGGCCAGGGCACGGCA
>hsa-mir-1244-2_5p
AUCUUAUUCCGAGCAUUCCAG
>hsa-mir-1244-3_5p
AUCUUAUUCCGAGCAUUCCAG
>hsa-mir-1972-2_5p
UGCCACCACACCUGGCUUAAAU
>hsa-mir-1302-9_5p
UAGCAUAAAUAUGUCCCAAGC
>hsa-mir-1302-10_5p
UAGCAUAAAUAUGUCCCAAGC
>hsa-mir-1302-11_5p
UAGCAUAAAUAUGUCCCAAGC
>hsa-mir-3118-6_5p
AAUUUUCAUAAUGCAAUCACAC
>hsa-mir-3609_5p
UUUUAUUCUCAUUUUCCUUUUC
>hsa-mir-3610_5p
UAACGGCAGCCAUCUUGUUUGUU
>hsa-mir-3611_5p
AGAAUUUCUUUUUCUUCACAAUU
>hsa-mir-3615_5p

AGACGCCGCGCGGGGCGGGGAUU
>hsa-mir-3618_5p
UGUGAUUUCCAAUAAUUGAGGC
>hsa-mir-23c_5p
CAGGUGUCACACAGUGAGUGG
>hsa-mir-3646_5p
AGGUUGGGUUCAUUUCAUUUUC
>hsa-mir-3649_5p
UGGAACAGGCACCUGUGUGUGC
>hsa-mir-3651_5p
CCUGUGAUUUAUGCAUGGAGGC
>hsa-mir-3653_5p
Not_Predicted
>hsa-mir-3654_5p
CAUGAGCUGCAAUCUCAUCAC
>hsa-mir-3656_5p
CGGCCAGCGGGACGGCAUCC
>hsa-mir-3657_5p
UCACCAAUAAUGGGACACUAA
>hsa-mir-3659_5p
CCCUUGUACACAACACACGUG
>hsa-mir-3660_5p
AAAAUGCUCUUCUGUCAUUGU
>hsa-mir-3662_5p
ACAGUUACACUUCUUACUCUCA
>hsa-mir-3669_5p
UACGGAAUAUAUAUACGGAAU
>hsa-mir-3670-1_5p
UCUAGACUGGUAUAGCUGCUUU
>hsa-mir-3671_5p
CUGCUGCUGUCACAUUUACAUG
>hsa-mir-3673_5p
UGGAAUGUAUAUACGGAAUAU
>hsa-mir-3684_5p
GGACCUGUACUAGGUUUAACA
>hsa-mir-3686_5p
AUUUACCUUCUCUUACAGAUCA
>hsa-mir-3687_5p
GCGCGUGCGCCCGAGCGCGGC
>hsa-mir-3907_5p
UGGGGUCCAGGCUGGACCAGG
>hsa-mir-3909_5p
UGGGGAGCAGGCUCCGGGGGACA
>hsa-mir-3910-1_5p
UCUUGGUUUUAUGCCUUUUA
>hsa-mir-3912_5p
CAUGUCCAUAUUAUGGGUUAGU
>hsa-mir-3914-1_5p
UCUCAUUUUCUGGUUCCUUCU
>hsa-mir-3914-2_5p
UCUCAUUUUCUGGUUCCUUCUACC
>hsa-mir-3917_5p
CCGGGUCUGUUGGUGCUCAGAGU
>hsa-mir-3919_5p
UGAGUCCUUUGUUCUCUACUA
>hsa-mir-3920_5p
CAGAGAGUUAAGAGAAUUAGUAC
>hsa-mir-3921_5p
AAGGCAUAUGGUACUCAAGAGA
>hsa-mir-3923_5p
CUAAUCCAAUAUUACUAGCUU
>hsa-mir-3910-2_5p
UCUUGGUUUUAUGCCUUUUA
>hsa-mir-3924_5p
UAGUAGUCAAAUAUGCAGAUCU
>hsa-mir-3928_5p
UGAAGCUCUAAGGUUCCGCCUG
>hsa-mir-3929_5p
AGUGGCUCACACCAGUAAUCCC
>hsa-mir-3935_5p
AGCUGAUGGUUGUAUCUAUGA
>hsa-mir-3936_5p
UGCUGUAGAUCCCUCAAAUCC
>hsa-mir-3937_5p
UCAGUUGCUACAGUUCCCUGUUG
>hsa-mir-3938_5p
AGAUUAUCUACAAGGGAAUUUU
>hsa-mir-3939_5p
UUCCUGUAUGUGGGCGUGCACG
>hsa-mir-3941_5p
UGAUGCUCAGUUGUGUGUAGAU
>hsa-mir-548z_5p
AAAAGUAAUUGAGAUUUUUGCU
>hsa-mir-548aa-1_5p
CAAAAGAAACUGUGGUUUUUGC
>hsa-mir-548aa-2_5p
CAAAAGAAACUGUGGUUUUUGC
>hsa-mir-378d-1_5p
ACUGUUUCUGUCCUUGUUCUUG
>hsa-mir-378e_5p
ACUCCAGUGUCCAGGCCAGGG
>hsa-mir-4418_5p
UUUUGCUCUGAGUGACCGUGGU
>hsa-mir-4419a_5p
UGUGCCUGUAGUCUUAGCUACU
>hsa-mir-378f_5p
UGGACUCCCAUAGUUUUCAGGCU
>hsa-mir-4420_5p
UUGGUAUGAACAUCUGUGUGUU
>hsa-mir-4421_5p
UCUCCUUUCUGCUGAGAGUUGA
>hsa-mir-548ac_5p
AAAAGUUAUUGUGGUUUUUGCU
>hsa-mir-4425_5p
GGUCCCAUUGAAUCCCAACAGC
>hsa-mir-4427_5p
UCUUGGGGCUUAUUUAGACA
>hsa-mir-4428_5p
UGCCAUGUUGCCUGCUCCUUA
>hsa-mir-548ad_5p
CAAAAGUAAUUGUGGUUUUUGA
>hsa-mir-4432_5p
GCAUCUUGCAGAGCCGUUCC
>hsa-mir-4436a_5p
CCACUUAUGCCUGCCCUGCCC
>hsa-mir-4437_5p
ACUUUGUGCAUUGGGUCCACA
>hsa-mir-548ae-1_5p

UGCCAUUAAGUUGCGGUUUUUG
>hsa-mir-548ae-2_5p
AAAAGUAAUUGUGGUUUUUGUC
>hsa-mir-4438_5p
UGUCUUUUCUAAGCCUGUGCC
>hsa-mir-4440_5p
AAGCAAGUGCAGUGGGGCUUGCU
>hsa-mir-4441_5p
CAGAGUCUCCUUCGUGUAC
>hsa-mir-4442_5p
GCGCCCUCCCUCUCUCCCCGG
>hsa-mir-4444-1_5p
UGGCCCCGCCUCUUCCUCUCGGU
>hsa-mir-4447_5p
UCUAGAGCAUGGUUUCUCAUCAU
>hsa-mir-4448_5p
ACCAAAAGACAAGAGUGCGAG
>hsa-mir-4449_5p
CCCUCGGCGGCCCGGGGGGCGG
>hsa-mir-4451_5p
UCUGUACCUCAGCUUUGCUCCCA
>hsa-mir-4452_5p
CACUUGAGGCCAAGAGUGCAAGG
>hsa-mir-4457_5p
CUCCAGUCAAUACCGUGUGAGU
>hsa-mir-4461_5p
UAGGUUAUGUACGUAGUCUAGG
>hsa-mir-4462_5p
UUCCCAGCUGCCCUAAGUCAGG
>hsa-mir-4463_5p
UGGUCACCACCUCCAGUUUCUG
>hsa-mir-548aj-1_5p
AAAAGUAAUUGCAGGUUAUGCC
>hsa-mir-4465_5p
CCCUGGCACGCUAUUUGAGGU
>hsa-mir-4468_5p
AGUCUUCUCCUGGGGCUUUGG
>hsa-mir-4469_5p
AGCGGCUCUAGGUGGGUUUGGC
>hsa-mir-4470_5p
UCGGCUUUCCAGUUUGUCUCG
>hsa-mir-4471_5p
AAACCUCUACUAAGUUUCCAUG
>hsa-mir-4472-1_5p
GACCCUUGCUCUCUCACUCUCC
>hsa-mir-4473_5p
CUUGUAAUGGAGAACACUAAGC
>hsa-mir-4475_5p
UCAAUGAGUGUGUGGUUCUAAA
>hsa-mir-4476_5p
CCUGUCCCUAAGUCCCUCCCAGC
>hsa-mir-4477a_5p
AAUCACAAAUGUCCUUAAUGGC
>hsa-mir-4477b_5p
AAUCACAAAUGUCCUUAAUAG
>hsa-mir-3689c_5p
UGUGAUAUCGUGGUUCCUGGGA
>hsa-mir-548x-2_5p
CACAAAAGUAAUUGUGGCUUUUG
>hsa-mir-4479_5p
AAGUCCGAGCGUGGCUGGCGCG
>hsa-mir-3155b_5p
CACUGCAGAGCCUGGGAAGGG
>hsa-mir-4480_5p
AGUUGACCUCCACAGGGCCACC
>hsa-mir-4483_5p
ACAACAUACUUAGUGCAUACC
>hsa-mir-4484_5p
GGGUUUCCUCUGCCUUUUUUU
>hsa-mir-4485_5p
AGAGGCACCGCCUGCCCAGUG
>hsa-mir-4487_5p
UGUCCUUCAGCCAGAGCUGG
>hsa-mir-548al_5p
UGCAAAAGUAAUUGCUGUUUU
>hsa-mir-4490_5p
UGCUCAAAUCUCUGGCCAAAGA
>hsa-mir-4491_5p
UGGUCACACCAGUCCACAUUAAC
>hsa-mir-4492_5p
UGCUUCUCCAGGCCCCGCGCGCG
>hsa-mir-4493_5p
AGAGAUGGGAAGGCCUUCCGG
>hsa-mir-4494_5p
CCCUGGUCAUCUGCAGUCUGAA
>hsa-mir-4496_5p
ACAUCAGCUCAUAUAAUCCUCGA
>hsa-mir-4419b_5p
AGUGGUGCAUGCUUAUAGUCC
>hsa-mir-4500_5p
AGGAGAGAAAGUACUGCCCAGA
>hsa-mir-4502_5p
UUUAGCAAGUUGUAAUCUUUUU
>hsa-mir-4504_5p
AGGUUCAUCUCUGUUGUCAUUUG
>hsa-mir-4506_5p
AUCAGACCAUCUGGGUUCAAG
>hsa-mir-2392_5p
AUCCAGCCAUUCCUCAGACCAG
>hsa-mir-4507_5p
UCUGGGCUGAGCCGAGC
>hsa-mir-4512_5p
GGCAAUAUAGUGAGACCUCGUC
>hsa-mir-4518_5p
UGCUGGGAUUGAUUAGUGAUGU
>hsa-mir-4522_5p
UGGGGGCCUCGCAGGGGGAGA
>hsa-mir-4526_5p
GGGCCCAGUCCCUGCUGUCAUG
>hsa-mir-4528_5p
ACAGAUCUUUAUAUAUAUGAU
>hsa-mir-4530_5p
CGACCGCACCCGCCCGAAGCU
>hsa-mir-4531_5p
GCCUAGGAGUCCUUGGUCAGUG
>hsa-mir-4534_5p
ACCCCCUUCCAGAGCCAAAAUC
>hsa-mir-4535_5p

ACUGGGUCCCAGUCUUCACAG
>hsa-mir-4540_5p
AAGCUGCAUGGACCAGGACUUGG
>hsa-mir-3960_5p
CGCGCCCCCGAUCGGGGCCGCC
>hsa-mir-3972_5p
GCUUGGGGUGGCAGUCCUGUGGG
>hsa-mir-3973_5p
AGGGUAGCUCUCUGUAUUGCUU
>hsa-mir-3975_5p
UGAGUGAUUGCUAUUUCAAAA
>hsa-mir-4635_5p
GUGGGUUCUGACCCCACUUGGAUC
>hsa-mir-4637_5p
ACUUGGAUCUGCAAUUAGUAUUU
>hsa-mir-4641_5p
GGGGGCAGGGGGCAGAGGGCAUC
>hsa-mir-4643_5p
AGCAUUUAUAAUCAUGUGUUCA
>hsa-mir-4644_5p
UCUGCCUCUUUCUCCAUCCACC
>hsa-mir-4658_5p
CCCUUCACUCAGAGCAUCUACAC
>hsa-mir-4662b_5p
UUAGCCAAUUGUCUAUCUUUAG
>hsa-mir-4669_5p
CCCUUCACUUCCUGGCCAUCC
>hsa-mir-4672_5p
CCUCUGUCCAGCUGUGUGGCC
>hsa-mir-4674_5p
CCCAGGCGCCCGCUCCCGACCC
>hsa-mir-4675_5p
UGCUGGUCAACCAUAGCCCUG
>hsa-mir-4683_5p
AGGCGGGCCUGGAGGUGCACC
>hsa-mir-1343_5p
UGGGGAGCGGCCCCCGGGUGGG
>hsa-mir-4688_5p
CAAGCUGUUUCGUGUUCCCUCC
>hsa-mir-4692_5p
CUUGAUACCCACACUGCCUGGG
>hsa-mir-4698_5p
GGGUCUUCCUCUACAUUUCCACC
>hsa-mir-3198-2_5p
ACUGUUCACCCAGCACUAGCA
>hsa-mir-4719_5p
UGUAUGUUAUAGAUUUGUGAUU
>hsa-mir-4721_5p
AUGGUCAAGCCAGGUUCCAUCA
>hsa-mir-4734_5p
CUCGGGCCCGACCGCGCCGGCC
>hsa-mir-4741_5p
CCGGCCGCCUCCGAGCCCGGC
>hsa-mir-4765_5p
AACGUAGCUAUCCACCACUCAG
>hsa-mir-4770_5p
GAGUUAUGGGGUCAUCUAUCC
>hsa-mir-4771-1_5p
UAAUUUUAGAUCUGGUCUGCUUC
>hsa-mir-4771-2_5p
UAAUUUUAGAUCUGGUCUGCUUC
>hsa-mir-4773-1_5p
CUUUCUAUGCUCCUGUUCUGCU
>hsa-mir-4773-2_5p
CUUUCUAUGCUCCUGUUCUGCU
>hsa-mir-4779_5p
GCUUUUACUGUUCCCUCCUAGA
>hsa-mir-4780_5p
AGGGGGUCAGGCUCAAGGACC
>hsa-mir-4785_5p
UGGGGACGCGGCGGCGCUGCU
>hsa-mir-4801_5p
AGGCUUGGUUUUCUUAUGUGUAA
>hsa-mir-5047_5p
ACGAGACACAGUGCAUAAAAA
>hsa-mir-5088_5p
AGGCGGGGCCGGGCCUGAGGG
>hsa-mir-5092_5p
UGCCAAAGCCAGUGGGGACUGG
>hsa-mir-5186_5p
CUGAUUUCUACCAACCUUUCCU
>hsa-mir-5188_5p
CUGGUUUCAAUGGGUACGAUUAU
>hsa-mir-5190_5p
UGGCUCCAGCCCUGUCACAUGG
>hsa-mir-5191_5p
CACUUCAUUCUUGCUGUCCUCU
>hsa-mir-5192_5p
ACCUGGAACCAUUUCUCCUGG
>hsa-mir-5193_5p
CUGGGAUGGGGGUUGGGGGGAG
>hsa-mir-5194_5p
AGGCCCAUUCUUUUCACUCAGGA
>hsa-mir-4444-2_5p
UGGCCCCGCCUCUUCCUCUCGGU
>hsa-mir-3670-2_5p
UCUAGACUGGUAUAGCUGCUUU
>hsa-mir-5100_5p
GGUAGGAGCGUGGCUUCUGGA
>hsa-mir-5572_5p
AGGCACUGCCCCUGCGACCAGCC
>hsa-mir-5680_5p
AGGUUAGCCCAGCAUUUCCCUUC
>hsa-mir-5681a_5p
AGAGUAUUGCCACCCUUUCU
>hsa-mir-5682_5p
CUUAUCCUGCAAGGUGCUGCA
>hsa-mir-5692c-1_5p
AUACCCACUGUGAUAUUAAGAGU
>hsa-mir-5692c-2_5p
ACACCAACUGUGAUAUUAGG
>hsa-mir-5687_5p
CUGACUCUGAAAUCUUCUAAAU
>hsa-mir-5688_5p
CCUUUUUACAGGAGUUUAUUAUG
>hsa-mir-5693_5p
AGUUAGUUCAUUUCAGUCUGUG
>hsa-mir-5695_5p

AUCUAGAUUCUUCUUGGCCUCU
>hsa-mir-5699_5p
CCCCAACAAGGAAGGACAAGAG
>hsa-mir-5702_5p
UGGGAUAUGUUGCUGAUCCAAC
>hsa-mir-5703_5p
GUCCCCUUCCUCGUCUUUU
>hsa-mir-5705_5p
AGGCCAUGAGCCCCGAAACACC
>hsa-mir-5739_5p
UAACUAUCAUUUCCAAGGUUG
>hsa-mir-6072_5p
UGGGGGCUGGUGCAGGGAUGGGC
>hsa-mir-6073_5p
AACUGAAAGUUGAUGAGUCACU
>hsa-mir-6074_5p
CACCUUCAGUCAACUGAUUUGC
>hsa-mir-6075_5p
CCACAUGCUCCUCCAGGCCUGC
>hsa-mir-6079_5p
UCCUAGACCUAGUAUCAGUGGCCA
>hsa-mir-6080_5p
UCUGGGAAUGCCGGUCUGGGGC
>hsa-mir-6081_5p
CACCAGGGCCUCUGCCCCGU
>hsa-mir-6083_5p
GUCUGGGAAGGUGGAAAGGGAG
>hsa-mir-6084_5p
UGGGCCGCAGGACCGGGCGCG
>hsa-mir-6085_5p
UGUGGGCCCAGCUUUACAUAGU
>hsa-mir-6089-1_5p
GGCCCGGCGUUCCCUCCCCUUCC
>hsa-mir-6090_5p
UGGGUCCGCGCGCCCUGGGCCG
>hsa-mir-6125_5p
GCUGCCACCUCCCCUACCGCUA
>hsa-mir-6128_5p
UAUAGGACUUCAGUCCAUGAU
>hsa-mir-6130_5p
AAAUGCAGGCAUCCCUUCUA
>hsa-mir-6131_5p
UCCCGCAUUCCCUCUGCUUUG
>hsa-mir-6133_5p
CAUGCCCUCUUCAUUGUUCUGCU
>hsa-mir-6715a_5p
ACAGGCACAGCCGGUUUGAGCA
>hsa-mir-6719_5p
GGAGGCUGAUGUCUUCAGAGC
>hsa-mir-6089-2_5p
GGCCCGGCGUUCCCUCCCCUUCC
>hsa-let-7c_3p
CUGUACAACCUUCUAGCUUUCC
>hsa-mir-196a-1_3p
CAACAACAUUAAACCACCCGAU
>hsa-mir-198_3p
UCCUUCUUCUCUAUAGAAUAAA
>hsa-mir-7-3_3p
CAACAAGUCACAGCCGGCCUCA
>hsa-mir-215_3p
UCUGUCAUUUCUUUAGGCCAAU
>hsa-mir-217_3p
CAUCAGUUCCUAAUGCAUUGCC
>hsa-mir-135a-2_3p
UGUAGGGAUGGAAGCCAUGAAA
>hsa-mir-134_3p
CUGUGGGCCACCUAGUCACCAA
>hsa-mir-190a_3p
ACUAUAUAUCAAACAUAUUCCU
>hsa-mir-194-1_3p
CCAGUGGAGAUGCUGUUACUUU
>hsa-mir-181b-2_3p
CUCACUGAUCAAUGAAUGCAAA
>hsa-mir-383_3p
CCACAGCACUGCCUGGUCAGA
>hsa-mir-325_3p
UUUAUUGAGGACCUCCUAUCAA
>hsa-mir-346_3p
AGGCAGGGGCUGGGCCUGCAGC
>hsa-mir-422a_3p
UCUCUGUCCCUGAGCCAAGCUU
>hsa-mir-449a_3p
CGGCUAACAUGCAACUGCUGUC
>hsa-mir-450a-1_3p
AUUGGGAACAUUUUGCAUGUAU
>hsa-mir-451a_3p
UAGUAAUGGUAAUGGUUCUC
>hsa-mir-484_3p
CCCGGGGGGUGACCCUGGCU
>hsa-mir-511-1_3p
AAUGUGUAGCAAAAGACAGAAU
>hsa-mir-511-2_3p
AAUGUGUAGCAAAAGACAGAAU
>hsa-mir-492_3p
CAGGAUUGUCCUGCAGAUCA
>hsa-mir-181d_3p
CCCACCGGGGGAUGAAUGUCA
>hsa-mir-498_3p
AAGCACCUCCAGAGCUUGAAGC
>hsa-mir-526a-1_3p
GAAAGCGCUUCCUUUUAGAGG
>hsa-mir-526a-2_3p
AACAUGCAUCCUUUCAGAGGGU
>hsa-mir-527_3p
GAAAGUGCUUCCCUUUGGUGAA
>hsa-mir-504_3p
AGGGAGUGCAGGGCAGGGUUUC
>hsa-mir-510_3p
UGAUUGAAACCUCUAAGAGUGG
>hsa-mir-553_3p
AUCUCGCUGUUUUAGACUGAGG
>hsa-mir-554_3p
GUGAUGGGUCAGGGUUCAUAUU
>hsa-mir-559_3p
UUUGGUGCAUAUUUACUUUAGG
>hsa-mir-564_3p
CCUCCGGGCGGCGCCGUGUCCGC
>hsa-mir-566_3p

CUGGGGCAGCAGAAUCGCUUGA
>hsa-mir-567_3p
UUGUACUGGAAGAACAUGCAAA
>hsa-mir-568_3p
UAGUGUAUAUAUUAUACAUGU
>hsa-mir-573_3p
UAAAGUUAUGUCGCUUGUCAGG
>hsa-mir-577_3p
GGUUUCAAUACUUUAUCUGCU
>hsa-mir-581_3p
UGAUCUAAAGAACACAAAGAAU
>hsa-mir-583_3p
UACUGGGACCUACCUCUUUGGU
>hsa-mir-586_3p
GCCCUAAAAAUACAAUGCAUAA
>hsa-mir-587_3p
UGACUCAUCACCAGUGGAAAGC
>hsa-mir-588_3p
UCUUACCCACCAUGGCCAAAA
>hsa-mir-592_3p
UCAUCACGUGGUGACGCAACAU
>hsa-mir-596_3p
CAUGGCAGCUGCUGCCCUUCGG
>hsa-mir-597_3p
AGUGGUUCUCUUGUGGCUCA
>hsa-mir-601_3p
CCCAGGGAUCCUGAAGUCCUUU
>hsa-mir-602_3p
CCGAGUGCGUCUCCUGUCAG
>hsa-mir-605_3p
AGAAGGCACUAUGAGAUUUAGA
>hsa-mir-608_3p
AGAGCUUCCAUCAAAGGCUGCC
>hsa-mir-609_3p
AGAGAUGAGGGCAACACCCUAG
>hsa-mir-610_3p
CCAGCACACAUUUAGCUCACA
>hsa-mir-612_3p
AGGGGCCCUCCCUCCAUGGCAG
>hsa-mir-617_3p
CACCUUCAAAUGGUAAGUCCAG
>hsa-mir-618_3p
UCAGGAGACAAGCAGGUUUACC
>hsa-mir-623_3p
CGAUGUACUCUGUAGAUGUCU
>hsa-mir-627_3p
CUCUUUUCUUUGAGACUCACUA
>hsa-mir-631_3p
UGAUGGACUGAGUCAGGGGCCA
>hsa-mir-635_3p
AUCAUUGUUUUGUGUCCAUUGA
>hsa-mir-638_3p
CGCGCCGUGCGCCGCCGGCGUAA
>hsa-mir-641_3p
GGUGACUGUCCUAUGUCUUUCC
>hsa-mir-647_3p
GGCAGGAGGGAGGGUCAGGCAG
>hsa-mir-648_3p
AAGUGCAGGACCUGGCACUUAGU
>hsa-mir-650_3p
CCUGGGCUCUGCUCCUCCUCA
>hsa-mir-651_3p
AAAGGAAAGUGUAUCCUAAAAG
>hsa-mir-663a_3p
UGGGAUCCCGCGGCCGUGUUUU
>hsa-mir-653_3p
UUCACUGGAGUUUGUUUCAAU
>hsa-mir-1296_3p
GAGUGGGGCUUCGACCCUAACC
>hsa-mir-1468_3p
GCAAAAUAAGCAAAUGGAAAA
>hsa-mir-1323_3p
AAAGUGCACCCAGUUUUGGGG
>hsa-mir-1283-1_3p
AAAGCGCUUCCCUUUUGAGGGU
>hsa-mir-378d-2_3p
UUCCCUGCUCUAAGUCCCAUUU
>hsa-mir-802_3p
AAGGAGAAUCUUUGUCACUUAG
>hsa-mir-670_3p
CUCAUAUUCAUUCAGGAGUGU
>hsa-mir-1298_3p
CAUCUGGGCAACUGACUGAACU
>hsa-mir-761_3p
AGUUUCACUUUGCUGCUCCUC
>hsa-mir-764_3p
AGGAGGCCAUAGUGGCAACUGU
>hsa-mir-759_3p
UAAAUGUUUGCACUGGCUGUUU
>hsa-mir-770_3p
UGGGCCUGAUGUGGUGCUGGGG
>hsa-mir-298_3p
AGGAACUAGCCUGCUGCUUUGC
>hsa-mir-891a_3p
AGUGGCACAUGUUUGUUGUGAG
>hsa-mir-890_3p
AACUAUUCCCUUUCUGAGUAGA
>hsa-mir-891b_3p
AAUGGCACAUGUUUGUUGUUAG
>hsa-mir-190b_3p
AACUAAAUGUCAAACAUAUUCU
>hsa-mir-216b_3p
CACACUUACCCGUAGAGAUUCU
>hsa-mir-921_3p
UCCAUGGGCCUGGAUCACUGG
>hsa-mir-924_3p
CAUCCAACCUAGAGUCUACAAC
>hsa-mir-934_3p
AGAGUCUCCAGUAAUGGACGGG
>hsa-mir-936_3p
UGAGAGACCUUGCUUCUACUUU
>hsa-mir-938_3p
CGUGGUACACCUUUAAGAACU
>hsa-mir-942_3p
CAUGGCCGAAACAGAGAAGUUA
>hsa-mir-297_3p
UAUGUAUUAUGUACUCAUAUAU
>hsa-mir-1179_3p

CAACCAAUAAGAGGAUGCCAU
>hsa-mir-1181_3p
GCCGCCAAGGCAAGAUGGUGG
>hsa-mir-1231_3p
UACCCUGUCUGUUCUUGCCACAG
>hsa-mir-1200_3p
UUGGUUCAGGAAUUUGUCAGG
>hsa-mir-1202_3p
CCACUCUGCUUAGCCAGCAGGU
>hsa-mir-1203_3p
CUCCAGAUUGUGGCGCUGGUGC
>hsa-mir-1204_3p
AGGUGAGGACGUGCCUCGUGGU
>hsa-mir-1205_3p
UGUCAACCCUGUUCUGGAGUCU
>hsa-mir-1206_3p
UUUUUGCAAGCUAGUGAACGCU
>hsa-mir-1208_3p
GCCCUCUGAUGAGUCACCACUG
>hsa-mir-548j_3p
CAAAAACUGCAUUACUUUUGC
>hsa-mir-1287_3p
CUCUAGCCACAGAUGCAGUGAU
>hsa-mir-1291_3p
UUGGUUUCAAGCAGAGGCCUAA
>hsa-mir-548k_3p
CAAAAACCGCAAUUAUUUUUGCU
>hsa-mir-1293_3p
ACAAAUCUCCGGACCACUUAGU
>hsa-mir-1294_3p
ACAACAGUGCCAACCUCACAGG
>hsa-mir-548l_3p
CAAAAACUGCAGUUACUUGUGC
>hsa-mir-1243_3p
UUACUUUGCUUUGGUAAUAAAUC
>hsa-mir-1246_3p
UGACCCAAAGGAAAUCAAUCCA
>hsa-mir-1248_3p
UAGCAGAGUACACACAAGAAGA
>hsa-mir-1250_3p
GGCCACAUUUUCCAGCCCAUUCA
>hsa-mir-1251_3p
CGCUUUGCUCAGCCAGUGUAG
>hsa-mir-1253_3p
CAGGCUGAUCUUCUCCCCUUU
>hsa-mir-1254-1_3p
CACUGUACUCCAGCCUAGGCAA
>hsa-mir-1255a_3p
UAUCUUCUUUGCUCAUCCUUG
>hsa-mir-1256_3p
CUAAAGAGAAGUCAAUGCAUGA
>hsa-mir-1257_3p
UGGCAUCACUGGCCCCAUCCUU
>hsa-mir-1260a_3p
CGGGGUCAGAGGGAGUGCCA
>hsa-mir-1261_3p
CUGAAACUUUCUCCAUAGCAG
>hsa-mir-1262_3p
UCCUUCUGGGAACUAAUUUUUG
>hsa-mir-1263_3p
UCAGUAUGCCAUGUUGCCAUAU
>hsa-mir-548n_3p
AAAACCCGCAAUUACUUUUGCA
>hsa-mir-548m_3p
AAAGCCACAAAUACCUUUGCA
>hsa-mir-1265_3p
AACAAUACUUGACCACAUUUUGA
>hsa-mir-1266_3p
CCCUGUUCUAUGCCCUGAGGGA
>hsa-mir-1267_3p
GGGAUUACAUUUCAACAUGA
>hsa-mir-1268a_3p
CCAGCUACUUUGGAGGCUGAG
>hsa-mir-1270-1_3p
CAGGCUUUUCUUUAUCUUCUAU
>hsa-mir-1272_3p
UAGAAAUGUAGGCUGCAGCUC
>hsa-mir-548h-1_3p
UAAAAACUGGAAUUACUUUUGC
>hsa-mir-548h-2_3p
CAAACACCACAAUUACUUUUGC
>hsa-mir-548h-3_3p
CAAAAACUGCAAUUACUUUUGC
>hsa-mir-1275_3p
UAUGCCAAACUUAUUUUCCCCAA
>hsa-mir-1276_3p
UGUCUCCACUGAGCACUUGGGC
>hsa-mir-302e_3p
UAAGAUGGAUGUAGUAAUAGCA
>hsa-mir-548i-1_3p
CAAAAAUAGCAAUUAUUUUUGU
>hsa-mir-548i-2_3p
CAAAAAUAGCAAUUAUUUUUGU
>hsa-mir-548i-3_3p
CAAAAAUAGCAAUUAUUUUUGU
>hsa-mir-548i-4_3p
CAAAAACCACAAAUAUUAUUGC
>hsa-mir-1279_3p
AAAGAAGAGUAUAAGAACUUCC
>hsa-mir-1282_3p
UGGGAGGUACCAGAGGGCA
>hsa-mir-1283-2_3p
AAAUCGCUUCCCUUUGGAGUGU
>hsa-mir-1284_3p
GAAAGCCCAUGUUUGUAUUGG
>hsa-mir-1252_3p
AAUGAGCUUAAUUUCCUUUUUU
>hsa-mir-1255b-1_3p
UACUCUUUGUGAAGAUGCUGUG
>hsa-mir-513b_3p
UAAAUGUCACCUUUUUGAGAGG
>hsa-mir-1469_3p
UGGGGCGAGCCAACGCCGGGG
>hsa-mir-1470_3p
CGCGGGACGCGCCGAGGUAGG
>hsa-mir-1471_3p
AGCUGGCUCUAAUUUGAGGGGC
>hsa-mir-103b-1_3p

CAAGGCAGCACUGUAAAGAAGC
>hsa-mir-103b-2_3p
CAAGGCAGCACUGUAAAGAAG
>hsa-mir-1908_3p
CCGGCCGCCGGCUCCGCCCCG
>hsa-mir-1910_3p
GGAGGCAGAAGCAGGAUGACA
>hsa-mir-2052_3p
UUACCAGCUAUCAAAACAA
>hsa-mir-2054_3p
UAUCAUUAAAAAAUGUAUUACA
>hsa-mir-2110_3p
UCUCACCGCGGUCUUUUCCUCC
>hsa-mir-548q_3p
UGAUUACUUUUUCACCAACCU
>hsa-mir-2278_3p
AGACAGCUUGCACUGACUCCA
>hsa-mir-2909_3p
AAGAUGGCUGUUGCAACUUAA
>hsa-mir-3115_3p
CCAACUAUUAGGCCUUUAUGU
>hsa-mir-3119-1_3p
CAUCAAAGUUAAAAGCCAUA
>hsa-mir-3119-2_3p
CAUCAAAGUUAAAAGCCAGA
>hsa-mir-3122_3p
GACCAUCAUCUUGUCCGAAGAG
>hsa-mir-3125_3p
UUCGAGAUCCCCGCCUUCCUCCU
>hsa-mir-3128_3p
UGAGAGUUUUUUACUUGCAAUAG
>hsa-mir-3131_3p
GGCCCUGGCCCCAGCUUCUUCUC
>hsa-mir-3132_3p
UUUCCCUUGAGCCCUCCCUCU
>hsa-mir-3133_3p
UGAGUUUUAAGAGUUCUUUAUA
>hsa-mir-3135a_3p
CUGCAGCCUUGACCUCCUGGGC
>hsa-mir-3137_3p
UCGUGCUCCUGGGCUACAAACC
>hsa-mir-3139_3p
CAGGUAUCGCAGGAGCUUUUG
>hsa-mir-3141_3p
CAUCAGCCUUCACUGGGACG
>hsa-mir-3143_3p
AGCAACUCUUUACAAUGUUUCU
>hsa-mir-1273c_3p
CAGAGUCUCGUUCUGUUGCCCA
>hsa-mir-3147_3p
CACCUCGCCCUUGUCCAACUCG
>hsa-mir-3148_3p
GCAUACAUCAGUUUUUUCCAAC
>hsa-mir-3151_3p
CAUCCCACCUGAUCCCACAGC
>hsa-mir-3159_3p
GGCUCACGCCUGUAAUCCCAGC
>hsa-mir-3161_3p
UGCUGGGCCUUUGUUUUUACC
>hsa-mir-3163_3p
CUUACUACCCCCAUUUUAUAGA
>hsa-mir-3164_3p
CCGUUUUGCUUGAAGUCGCAGU
>hsa-mir-3165_3p
UGAGGUCACAUUGUAUCCACCU
>hsa-mir-1260b_3p
UGGUGAUAGUCUGGUGGGGGCG
>hsa-mir-3168_3p
CUGUGUGACCCUGGGCCAGUG
>hsa-mir-3169_3p
GUGUGCCAAGCAUAGUCCUGUG
>hsa-mir-3170_3p
UGCCUGUCUUAGAACCCCUAUG
>hsa-mir-3171_3p
UAUAUAGAUUCCAUAAAUCUAU
>hsa-mir-1193_3p
UAGGUCACCCGUUUGACUAUCC
>hsa-mir-3174_3p
UACUGGAUCUGCAUUUUAAUUC
>hsa-mir-3175_3p
UCGGGCGCUUUCUCCUCCCCCU
>hsa-mir-3178_3p
UCUCCCGUGCCCACGCCCCAAA
>hsa-mir-548w_3p
CAAAACCCACAAUUACUUUUGC
>hsa-mir-3181_3p
CGAGCCGGCCGGGCCCGCGGGU
>hsa-mir-3182_3p
UGCUUGGGUUGGAUCAUAGAGCAG
>hsa-mir-3183_3p
CGGCGCCUCCUCGAGGGAGGAGA
>hsa-mir-3185_3p
AGAGCCAUCCGCCUUCUGUCCA
>hsa-mir-3192_3p
UUCCUCUGAUCGCCCUCUCAG
>hsa-mir-3193_3p
UUACCCAGCUCCUGAGCAGG
>hsa-mir-3195_3p
GCCGGGGCGGGGGCGGGGGCUG
>hsa-mir-3196_3p
GGCCCCAUUCUGCUUCUCUCCC
>hsa-mir-3156-3_3p
CUCCCACUUCCUGAUCUUUCU
>hsa-mir-3197_3p
CCUUCCCGAUCCACCGCUCUC
>hsa-mir-3199-1_3p
UUUCUCCUAAGGCAGUCCCUGG
>hsa-mir-3199-2_3p
UUUCUCCUAAGGCAGUCCCUG
>hsa-mir-3201_3p
CCUCUUAUUUUUACAUGCCC
>hsa-mir-3202-1_3p
UAAAGCUCUUCUCCCUUCCAUA
>hsa-mir-3202-2_3p
UAAAGCUCUUCUCCCUUCCAUA
>hsa-mir-1273d_3p
UGCACUUCAGCCUGGGUGACAA
>hsa-mir-4295_3p

UCAAGGCUAAGAAACUAGACUG
>hsa-mir-4296_3p
UGGGGACUGUGUCCAUGUCU
>hsa-mir-4297_3p
CGGCAGGGCCAGGACGGGUCGC
>hsa-mir-378c_3p
AACUCUGACUUUGAAGGUGGUGA
>hsa-mir-4294_3p
GCCCAAGGGUGCAUGUGUCU
>hsa-mir-4301_3p
GUGGAGGGUGGCAGGUGCAGC
>hsa-mir-4298_3p
CAGAAACCAAACUGUCAAAAGU
>hsa-mir-4304_3p
CUCUGUGACUGCUGCCAUCCU
>hsa-mir-4302_3p
GGCUGAGUUUACUUAAGGU
>hsa-mir-4303_3p
UUCUUUAGCUUAGGAGCUAACC
>hsa-mir-4305_3p
CUGUGAGGGAAAUUCUCUGU
>hsa-mir-4309_3p
UUGUAGGGUCUGCGGUUUGAAG
>hsa-mir-4311_3p
CAGGUUACAGGUUCGAUCUUU
>hsa-mir-4313_3p
GUGGGGGAAUCAGGGGUGUAA
>hsa-mir-4315-1_3p
UCUGGUGCAGAACUACAGCGG
>hsa-mir-4316_3p
CCAGCCCAGCCCCAAUCCCACCA
>hsa-mir-4314_3p
UUCUGCCCCAGGGCCCAGAGU
>hsa-mir-4319_3p
GUGCUCAGCUCAUGGGGCUA
>hsa-mir-4317_3p
UAGCUCUCUUGAUAAAAUGUUU
>hsa-mir-4256_3p
AUUGAUUAGGUCUGAUGAUCCA
>hsa-mir-4258_3p
CUGGGCUUGGUUUGGGGGCGG
>hsa-mir-4260_3p
CCCACACCCCAGCUUGUCACAC
>hsa-mir-4255_3p
CCAUUUUUAGGGCAAAGAGGCA
>hsa-mir-4325_3p
AAGGAUGGAGAGAAGGCAGAUC
>hsa-mir-4326_3p
CUGGGUGGAUGGAGCAGGUC
>hsa-mir-4327_3p
AGGGAGUUCUCAUCAAGCCUUU
>hsa-mir-4267_3p
CCCAGCCUCUGUCAUCCCUGCAU
>hsa-mir-4269_3p
GGAAGCCACUCUGUCAGGCCUG
>hsa-mir-4264_3p
UGACAGGUACUGGGUAAGACU
>hsa-mir-4270_3p
GCCUUCCCCUGCUGGGAAGA
>hsa-mir-4273_3p
AUGCUUCUUCACAAUGGUCACA
>hsa-mir-4276_3p
UAAAUAGAGCUACUGUGUCUGA
>hsa-mir-4275_3p
AUAAAAAAGUGAUAAUGGGAA
>hsa-mir-4277_3p
CAGUGCCCUGCUCAGCUCAAGU
>hsa-mir-4280_3p
UAUGUUCAAGACUGAAUGACA
>hsa-mir-4284_3p
AGAGGGGGUAGUUAGGAGCUUU
>hsa-mir-4286_3p
UACCAUGACUUAAGUGUGGUGG
>hsa-mir-4287_3p
UGUGGUCCCUACUGGGGAGACC
>hsa-mir-4289_3p
CUGGGCCCUGUCUGCAGAGCC
>hsa-mir-4291_3p
GCUGUUCUGCUGUGGCUGCAG
>hsa-mir-500b_3p
AGUGCACCCAGGCAAGGAUUCU
>hsa-mir-1270-2_3p
CAGGCUUUUCUUUAUCUUCUAU
>hsa-mir-4315-2_3p
UCUGGUGCAGAACUACAGCGG
>hsa-mir-3612_3p
CAGUUCACUAGAGGCGUCCUGA
>hsa-mir-3621_3p
CCACCUGACGCCGCGCCUUUGU
>hsa-mir-3648_3p
CUCGAGGGGUCCCCGUGGCGU
>hsa-mir-3650_3p
GCUCUGUCUGGCACAUUUCUGA
>hsa-mir-3652_3p
CUGGGCCUCUGCUGCGUCCUG
>hsa-mir-3655_3p
CAAAAUGCCGGAGCGAGAUAGU
>hsa-mir-3658_3p
UGAUUUUUUUUUUCUUUUUGUA
>hsa-mir-1273e_3p
CAGCCUGGGUGACACAGCGAGA
>hsa-mir-3661_3p
GCUGCUCGAUCCACUGGUCC
>hsa-mir-3665_3p
ACUCCGCAGCUGCUCGUUCUG
>hsa-mir-3666_3p
AGCGUUUCACACUGCCUGGU
>hsa-mir-3668_3p
UUUGAUCAAUCUCUGCAAUUUU
>hsa-mir-3672_3p
GAUGUUUUAUGAGUCUCAUGA
>hsa-mir-3674_3p
UCCUUUCAAGUUUUUGCAUUUC
>hsa-mir-3683_3p
AUGCUACGAACAAUAUCACAGA
>hsa-mir-3685_3p
UUGGGGGGAUGGGCAAAGUAC
>hsa-mir-3690-1_3p

CAUAUCUACCUGGACCCAGUGU
>hsa-mir-3713_3p
UAUCCCCAGAUGGAUACCAA
>hsa-mir-3714_3p
GGAAGACACCGCUGCCACCUC
>hsa-mir-3180-4_3p
CCUCCGGAUGCCAGUCCCUCAU
>hsa-mir-3180-5_3p
CCUCCGGAUGCCAGUCCCUCAU
>hsa-mir-3908_3p
CAGAGUCUCCCUCUGUCGCCAGG
>hsa-mir-3911_3p
CCUGCGCUUCUCGAUUCCCAGA
>hsa-mir-3915_3p
UAAGACCAUCCUUUCCUCAAU
>hsa-mir-3916_3p
UUCAGGGGAUGUGUCUCCUCUU
>hsa-mir-3918_3p
UCUCCAGCUGGGACCCUGCAC
>hsa-mir-3926-1_3p
UCUGCCUGCUUUUUGGCCAGC
>hsa-mir-3926-2_3p
UCUGCCUGCUUUUUGGCCAGC
>hsa-mir-3943_3p
CUAAGUAAAGUGGGGGGUGGG
>hsa-mir-3945_3p
UACAGCCUCUUAUGCUUUCC
>hsa-mir-1254-2_3p
CCUGUACUCUAGCCUGGGCA
>hsa-mir-1268b_3p
UAAUUCCAGCUAGUUGGGA
>hsa-mir-548h-5_3p
UAAAAACCACGGUUGCUUUUGC
>hsa-mir-548ab_3p
CAAAACCCGCAAUUAGUUUUGC
>hsa-mir-4417_3p
CCAGCAUCCAGGGCUCACCUAC
>hsa-mir-4422_3p
UGGGCCCUUCUUGAUGCUCUUG
>hsa-mir-378g_3p
UGGCUCCAGCCCAGCUC
>hsa-mir-4424_3p
UUAGUCCAUUUCAAGUUAACU
>hsa-mir-4426_3p
AGGAGUCUACUCUUCAUCUUG
>hsa-mir-4429_3p
UUUGUCUCUCCAACUCAGACU
>hsa-mir-4430_3p
CACUGCACUCCAACCUGGUGA
>hsa-mir-4431_3p
UUUCUAGUUGUCAGAGUCAUUA
>hsa-mir-4434_3p
UUUCAACUUUUCCUACAGUGU
>hsa-mir-4435-1_3p
AGUGUGACUCAGCAGGCCAACA
>hsa-mir-4435-2_3p
AGUGUGACUCAGCAGGCCAACA
>hsa-mir-4439_3p
AAGGUAUCAGUUUACCAGGCCA
>hsa-mir-4443_3p
UAUCCCUUUCUAGCCCUGAGCA
>hsa-mir-548ag-1_3p
CAAAUAUUACAUUUACUUUUGC
>hsa-mir-548ag-2_3p
CAAGAACCUCAAUUACCUUUGC
>hsa-mir-4450_3p
CACCAUCUCCCCUGGUCCCUUGG
>hsa-mir-4453_3p
AGGAGGCCCAGGCCGCGUCUUC
>hsa-mir-4454_3p
UGUCCGUGUGAAGAGACCACCA
>hsa-mir-4455_3p
GAAGGACAGCCAAAUUCUUCA
>hsa-mir-4456_3p
UGGGAGGAAGUUAGGGUU
>hsa-mir-4458_3p
UUUUAGUUACACUCUGCUGUGG
>hsa-mir-4459_3p
UGGCACUGACUCCAGCCUGGGG
>hsa-mir-4460_3p
GGUAAAUUCACAACCACUGUGG
>hsa-mir-378h_3p
UAGCAGCAAUCUGAUCUUGAGC
>hsa-mir-3135b_3p
UCACUGCAGCCUCGAACUCC
>hsa-mir-4464_3p
UAUCCAAACCUUACUAAUUCA
>hsa-mir-548ai_3p
AAAAAAAAUCACAAUUACUUUU
>hsa-mir-4466_3p
CCGGCCCGGCCCGGCCCGGCGA
>hsa-mir-4467_3p
CCCCUGGGCCGCCGCCUCCCU
>hsa-mir-4472-2_3p
UCUUGCUCCGUCGCCCAGGCCG
>hsa-mir-4478_3p
AGCCUCAUCCCCUGCAGCCCUG
>hsa-mir-3689d-1_3p
UGUGAUCCUGUUCUUCCUG
>hsa-mir-3689d-2_3p
UGUGAUCCUGUUCUUCCUGAGC
>hsa-mir-3689e_3p
CUGGGAGGUGUGAUCCCGUGC
>hsa-mir-3689f_3p
CUGGGAGGUGUGAUCCCACACU
>hsa-mir-548ak_3p
AAACCGCAAUUACUUUUGCAG
>hsa-mir-4481_3p
CUAGCACAUGAGCACGCUC
>hsa-mir-4486_3p
UAGAUGCUUGCUCUUGCCAUUG
>hsa-mir-4488_3p
CGCCUUGGCCCCGCCCCGCCC
>hsa-mir-4489_3p
UCCUGCCUGACCCUGUCCCA
>hsa-mir-4495_3p
AGCAAAAAGCUUAUUUACAUUU
>hsa-mir-4497_3p

CCCGGCGCCCGUCCGCCCGCGG
>hsa-mir-4498_3p
AGCAGCCCCUGCCUUGGAUCUC
>hsa-mir-4499_3p
UAACUCCUUGUCUCAGUCUGUU
>hsa-mir-4501_3p
UUUCAUCAGAUGUCACAUUUU
>hsa-mir-4503_3p
GUUUCUAUUUCCUGCUUAAAUA
>hsa-mir-4505_3p
UCCUCCAUGUCGGCCCGCCUUG
>hsa-mir-4508_3p
CCUGCGCCGGCAGCUGCAAGG
>hsa-mir-4509-1_3p
AACCUUCUGUAUCCUUUAUUUU
>hsa-mir-4509-2_3p
AACCUUCUGUAUCCUUUAUUUU
>hsa-mir-4509-3_3p
AACCUUCUGUAUCCUUUAUUUU
>hsa-mir-4510_3p
CUACAAUCUUUUCUCACAACA
>hsa-mir-4511_3p
GGUAAAUGCAAUAGUUCUUCUU
>hsa-mir-4513_3p
CGGCCCCCAGAUUUCUGGUCUCC
>hsa-mir-4514_3p
UACCUCGUCUCUUGCCUGUUUUAG
>hsa-mir-4515_3p
UGGGGAGCUGGUCCUAGCUCA
>hsa-mir-4516_3p
CACGGCUCUGCCCACGUCUCCC
>hsa-mir-4517_3p
UGUGUUUGGGUGGUGUGGCUG
>hsa-mir-4519_3p
GCGGCCUGCAGUAAGCGGGUA
>hsa-mir-4521_3p
AAACUAGGAUUUCUCUUGUUAC
>hsa-mir-1269b_3p
AGAUGGCUUAUCAUGGGACCUCU
>hsa-mir-4523_3p
CGGCCGAGGCCCGGGCCGGUUC
>hsa-mir-4525_3p
UCAGCGUGCACUUCCCCACCCUG
>hsa-mir-4527_3p
CAUCAGCUCUGUGCUGCCUAC
>hsa-mir-4532_3p
CCUGGUAUCCUGGGUGU
>hsa-mir-4533_3p
GUCCACUUCCCUUUCUCUCUCU
>hsa-mir-378i_3p
UUCCCACUCUUGGGCCCUGGGC
>hsa-mir-1587_3p
UGGACUCACCUGUGACCAGC
>hsa-mir-548an_3p
CAAAAACCGCAAUUCCUUUUGC
>hsa-mir-4537_3p
GCUGAGCUGGGCUGAGCUGAGC
>hsa-mir-4538_3p
AGCCAGGCUGAUCUGGGCUGAG
>hsa-mir-3974_3p
UGACAAAUUUGACUACAGCCU
>hsa-mir-3976_3p
AUGACAUGGGAUUUGGCUGUU
>hsa-mir-3977_3p
CUUUAAUUUGUUAUGUGUUGGCA
>hsa-mir-3978_3p
CUUGGGCAUCGUUUUCUUUUGA
>hsa-mir-4634_3p
UUGGGCGGCCGCGUUUCCCCUCC
>hsa-mir-4636_3p
UAAAGGCUUCAAGCACGAGUUCU
>hsa-mir-4642_3p
CAAAGCCACUCAGUGAUGAUGC
>hsa-mir-4647_3p
CCCAGCACCACCACCUCCUAU
>hsa-mir-4648_3p
CCCCUGCUCUGUUCCCACAG
>hsa-mir-4651_3p
UUUGCCGGGCGCCUCAGUUCA
>hsa-mir-4654_3p
CAGUCUCCUUUUCCCUCAUCAU
>hsa-mir-4656_3p
GCCUCCUGCUUCCUGGGCUCAG
>hsa-mir-4657_3p
UGCCAAGAACACUACCAUAU
>hsa-mir-4660_3p
UCCAUCUCCCCCAGGGCCUGG
>hsa-mir-4663_3p
UUCCUGGAGCUCAGGCCCUUGC
>hsa-mir-4673_3p
CUGACCCGGCCCCUCUUGCGG
>hsa-mir-4678_3p
AAGAUUCUGAGCAAUAACCUAU
>hsa-mir-4679-1_3p
CAAAGAAUCUCUAUCACAGAAA
>hsa-mir-4679-2_3p
CAAAGAAUCUCUAUCACAGAAA
>hsa-mir-4681_3p
CAGGUGCAGGCUGCAGACCUGU
>hsa-mir-4682_3p
UCCAGAGCUCCAAGGCUCAGUGC
>hsa-mir-4686_3p
CACCCUGGGCCCAGCAGGAGCC
>hsa-mir-4689_3p
CCAUGCCAUGUGUCCUCAUGG
>hsa-mir-4696_3p
UGACAAUGUCCAUUUUGCAGU
>hsa-mir-4705_3p
AGCAAUUACCAAGUGAUUGGUU
>hsa-mir-4706_3p
CAGCCCACUCCUGUCCUGGGCU
>hsa-mir-4710_3p
AGCAGCUCUCGCCUCUUCGUC
>hsa-mir-4718_3p
CUUGGCUUCAGUUACUAGC
>hsa-mir-451b_3p
UGGUAACGGUUUCCUUGCCAUU
>hsa-mir-4729_3p

GCGUCCCAGCAGAUAAAUGAGG
>hsa-mir-4730_3p
GCUGGUGUGUGCUGCUCCACAG
>hsa-mir-4736_3p
UGGCUGCUUGCCUGCCU
>hsa-mir-4737_3p
ACGGUGCCUCACAGCCACACAG
>hsa-mir-4739_3p
AGCCUCUCCCUUCCUCCCCUCC
>hsa-mir-4744_3p
UAGUAUGUCUAGUCUUUAGGUU
>hsa-mir-4748_3p
CAGACCCUACCCAACCCCACG
>hsa-mir-4751_3p
CUUCUGGGGGCUGGUCUUCAG
>hsa-mir-4752_3p
CAUGUUCUUCAGAUGGACAAGG
>hsa-mir-4754_3p
UCCGCAGGUCCAGGUUGCCGUG
>hsa-mir-4759_3p
AAUUCCAACAUCUAGUCCUAAA
>hsa-mir-4767_3p
CCGGGGGCAGAGCGCGCGGGAG
>hsa-mir-4775_3p
UGACUGAAACAAAAAAUUAAAA
>hsa-mir-4784_3p
CGUCCCUGCUCAUCCUCUCCGC
>hsa-mir-4788_3p
CUCCCUUAGUUGGUCCCUAAUC
>hsa-mir-4791_3p
UAUGUGCAGUCAAUGUCCAGU
>hsa-mir-4792_3p
UGGGGCCGCGCACAUCUCUGC
>hsa-mir-4794_3p
UAGUCUCAUGAGAUAGCCAGAUG
>hsa-mir-4803_3p
CAACCCACACUAUGAUGUUAAA
>hsa-mir-5087_3p
AGUCGCAAGCAUAAGAAAGAGA
>hsa-mir-5090_3p
UAAGCCUUCUGCCCCCAACUCC
>hsa-mir-5091_3p
UCACCGGCAGGGGUCUGGAGUC
>hsa-mir-5094_3p
UGGUAGGUACAGUGGGCUCAC
>hsa-mir-5095_3p
CGGUGGCUCACGCCUGUAAUC
>hsa-mir-1273f_3p
CUGCACCCCCAGCCUGGGCCA
>hsa-mir-5096_3p
UGACCUCAGGUGAUCCAUCCAC
>hsa-mir-5189_3p
UGCCAACCGUCAGAGCCCAGA
>hsa-mir-548aw_3p
CCGCGAUGACUUUUGCAUCAAC
>hsa-mir-5683_3p
AGUCAGGAUCUGCAUUUGAAUA
>hsa-mir-5684_3p
UGUUGCCCAGGCUGGAGUCCA
>hsa-mir-548ax_3p
CAAAAACCGUAAUUACUUUUGU
>hsa-mir-5685_3p
CGUGAUAACUGCAGGGCUGUGA
>hsa-mir-5686_3p
UGUAUUGUAUCGUAUCGUAUCG
>hsa-mir-5681b_3p
UAGAAAGGGUGGCAAUACUCU
>hsa-mir-5689_3p
UAGGACUACAGGUGUGUGCUA
>hsa-mir-5690_3p
UAAUAGAGGUAAUAGUUGAAA
>hsa-mir-5691_3p
CUGCUUGGUGUUCAGAGCUUGU
>hsa-mir-5692a-1_3p
ACACCCUGUGAUAUUAUUUGUA
>hsa-mir-5692a-2_3p
ACACCCUGUGAUAUUAUUUGUA
>hsa-mir-4666b_3p
GAAUUACAAUUUGACAUGCAAUU
>hsa-mir-5694_3p
CUGAGAAGUCCCAUGAUCCGC
>hsa-mir-5696_3p
CGUCAGACUACCUAAAUGAGCAC
>hsa-mir-5697_3p
CUUUAUCAUGAAACGCUUGAGG
>hsa-mir-5698_3p
ACAAUCACUGUACUCCCCAGG
>hsa-mir-5700_3p
AUAAUUUAAUGCAUUUAUUGA
>hsa-mir-5701-1_3p
UCAGAACAUGAAAAUAACGUCCA
>hsa-mir-5692b_3p
UACACCCAUGUGAUAUUUGAAG
>hsa-mir-5704_3p
AUAACAGGAUGAUGGCCUAAAC
>hsa-mir-5706_3p
CUUCAGCAUGUUUUCCAGAGG
>hsa-mir-5707_3p
AUGUACAGCUUUCAAACAUGCU
>hsa-mir-5708_3p
UCUUGGCCAGGCACAGUGGCUC
>hsa-mir-5701-2_3p
UCAGAACAUGAAAAUAACGUCCA
>hsa-mir-5787_3p
CUCGGCUCCCGCGCCGCACCC
>hsa-mir-6068_3p
CUGCUGGCGCAGGCUCGGCC
>hsa-mir-6069_3p
AGGGUGGAGGGUCACUCCUUA
>hsa-mir-6070_3p
UGAGCUCUUGUUGAUUGCAGUG
>hsa-mir-6071_3p
UCAGAACCCCCGCCACCACAGA
>hsa-mir-6076_3p
CCCUUUCACCCUCCUGAGUUUGG
>hsa-mir-6077-1_3p
UUAAGGCUGACGCUCCCUAUU
>hsa-mir-6078_3p

UCAGCUGGUUUUGAGUGAGAAG
>hsa-mir-6082_3p
AUUGACUGGGCUAUAUUUGUC
>hsa-mir-6086_3p
UUUGCCUUGUUUUUCUUUUU
>hsa-mir-6087_3p
GGCUCUCGCUUCUGGCGCCAAG
>hsa-mir-6088_3p
UAUUGCCUACGCUGAUCUCA
>hsa-mir-6124_3p
UGCCACUCCUGCCCAGUGCCUC
>hsa-mir-6126_3p
UCUGCCCACCCCACACCCUGCCU
>hsa-mir-6127_3p
UCCUCCUCCUCCUCCUCCUUC
>hsa-mir-6129_3p
UUGUCCAAGUUUCCCUUUGAA
>hsa-mir-6132_3p
CUGCUCCCAGUCCUGCCCCUGC
>hsa-mir-6717_3p
UUUCCUCAUCCUGCCAGGCCACC
>hsa-mir-6718_3p
AUAAGCCUUUUGGCCACUAGG
>hsa-mir-6721_3p
UGACCUGCUUUAACCCUUCCCCA
>hsa-mir-6723_3p
UGGGAAGAAAGUUAGAUUUACG
>hsa-mir-6724_3p
UCCCGAGGCCCGAGCCGCGACC
>hsa-mir-3690-2_3p
CAUAUCUACCUGGACCCAGUGU
>hsa-mir-6077-2_3p
UUAAGGCUGACGCUCCCUAUU
>mmu-mir-207_5p
UGAGGGGCUGCGGGAGGAGCCGG
>mmu-mir-762_5p
UCUCGGCCCGCACGGUCCGGCC
>mmu-mir-3475_5p
CAAAUCAUGUACCCCCACAG
>mmu-mir-678_5p
AGCUGUGCUCCAAUAUGAGAGA
>mmu-mir-682_5p
UCUGGCACUGUGGUUCCUGCA
>mmu-mir-683-1_5p
AGGCUGCAGUGGACCCAGGCU
>mmu-mir-684-1_5p
UUAAGUAGGGAUAAAUUACUCU
>mmu-mir-684-2_5p
UUAAGUAGGGAUAAAUUACUCU
>mmu-mir-688_5p
AAGAAAAGUAGGGGCUUGCUUG
>mmu-mir-690_5p
UGUGGAGCUAAUUGGCUGUAUU
>mmu-mir-691_5p
UUUUGCUUUCUUCCUUGGGUCU
>mmu-mir-692-1_5p
AGACUGGCGCGCCCCAGGGAUCU
>mmu-mir-692-2_5p
AGACUGGCGCGCCCCAGGGAUCU
>mmu-mir-694_5p
UCAGGCAUCGCUUUCAACCC
>mmu-mir-697_5p
UUGACAGGUCUCAGAGGUGACU
>mmu-mir-704_5p
UGGGAGCUAGAGGAUGUGGUCA
>mmu-mir-709_5p
UGUCCCGUUUCUCUGCUUCU
>mmu-mir-711_5p
AAUCUCUUCUUAGGGUGCUUC
>mmu-mir-713_5p
UUAGUGAGACUUGAUUGACAUG
>mmu-mir-718_5p
AGGCCGCGGAGGGCAAGAUGG
>mmu-mir-804_5p
AGGUUACAACUUCCCCAGUAGA
>mmu-mir-343_5p
UGGGAUAGAGUGGGUGUGGCGGG
>mmu-mir-453_5p
CAGGAGUGCUGUGAGAAGUG
>mmu-mir-466g_5p
UGUGUGCAUGUGGAUGUAUGU
>mmu-mir-1187_5p
UUACACACACACACACACACA
>mmu-mir-669j_5p
GUGCAUGUGUGUAUAGUUGUGU
>mmu-mir-1190_5p
CGUGGGAAGGUCUCUGCUGGC
>mmu-mir-466j_5p
UUGUGCAUGUGUGUAUGUGUGC
>mmu-mir-467g_5p
AUAUGUGUGUGUGUAUAUAUA
>mmu-mir-1902_5p
GGAGUGUUUGCUGUAUAAUUGG
>mmu-mir-1905_5p
UGCUGCUGGAUGCGUUUGAUGGU
>mmu-mir-1895_5p
UUUCCUCUUCUUCUUGGCCGGG
>mmu-mir-1900_5p
AAGCUAGAAGAGGGCGGAGCCU
>mmu-mir-1892_5p
UUCUACUCUUGACCAAAGUUU
>mmu-mir-1896_5p
CUCAUUACAGUGAUGUCUUU
>mmu-mir-1907_5p
AGCCCCACUCCCCCUCGCUGUC
>mmu-mir-1893_5p
AGGUAUCUGCUGCGCCUGAGAUG
>mmu-mir-1901_5p
UUCCCUGAGUGAACGAGUUGAG
>mmu-mir-1927_5p
AGAUCCCUAGAAACCAGAGUUG
>mmu-mir-1928_5p
AGGAUAGAGCUUUGCGCAUUG
>mmu-mir-1931_5p
AGCCAUCUCCUAGGCCAGAA
>mmu-mir-1932_5p
AGCCUGGCCCUGAGUCUCCGACCC
>mmu-mir-1936_5p

CCAGUCACACAGAGGACUUUAG
>mmu-mir-1942_5p
AGGCCUAUUUAAUGUUAGACA
>mmu-mir-1945_5p
CCCGUCAGCCCCCGAGAAAAC
>mmu-mir-1949_5p
GCUGGUUGGCAUUCUGGGCCU
>mmu-mir-1951_5p
AACCACCUCUCUUCACUUACUU
>mmu-mir-1958_5p
AGAACUUACUGCUUCCACUUUC
>mmu-mir-1961_5p
AGUUCAACACGCCCUCCCCUCU
>mmu-mir-1962_5p
UGGUCCCAUCUGUCUGCCUCUCU
>mmu-mir-1965_5p
CGCCACCAUGCCUGGCUCUUUU
>mmu-mir-1967_5p
UCUUUCUCUCUUUGCUCCUUU
>mmu-mir-1946b_5p
AGGCGUGCGCCACCACUGCCCA
>mmu-mir-1983_5p
AAAGCAUGCUCCAGUGGCGCA
>mmu-mir-683-2_5p
AGGCUGCAGUGGACCCAGGCU
>mmu-mir-2136_5p
CCAGUCAGGAGUCAGUUAGGA
>mmu-mir-2137_5p
ACCUCCUUCCUGCUGCCUU
>mmu-mir-2139_5p
AGCAGAGGGCCAGGACUGGCAUU
>mmu-mir-432_5p
UAGCUCUUGCAUUUCCUGGUGG
>mmu-mir-599_5p
UAUUUGAUAAGAUGACAUAGGA
>mmu-mir-2861_5p
UCCGGCUCCCCCUGGCCUCCC
>mmu-mir-3472_5p
UUUCCAGCUUCUGGCUAUUAUA
>mmu-mir-3473a_5p
CCUGUUGAGCCAUCUCACCAG
>mmu-mir-3960_5p
UCCUGCGCCCCCGAUCGGGGCC
>mmu-mir-3961_5p
AGAGGACCAAAUGCACUCAGAGC
>mmu-mir-28c_5p
ACAAAGACAAAUGAGAUAUGA
>mmu-mir-3962_5p
UGAGAAAUGUACUCUGCCACG
>mmu-mir-3964_5p
GGCCUGCUUUCCAAGUUAUGU
>mmu-mir-3965_5p
CAGAGAGCUGCAGCUGAGUGC
>mmu-mir-378b_5p
UCCUGGGCUAUCCCAGUCCAGG
>mmu-mir-101c_5p
UACUGCACAGUCCUGUGAUGA
>mmu-mir-3967_5p
UGCACCUGACUCAGGCAGCAA
>mmu-mir-3968_5p
AGCGUGUGGUGGUAGGAUCCGU
>mmu-mir-3971_5p
AGGUGGAAUGGGAGGUGGCAGG
>mmu-mir-3473b_5p
UGAGCCAUCUCUCCAGCCCAA
>mmu-mir-5097_5p
CGGACAGAUGGGCAUGGAGUCG
>mmu-mir-5099_5p
AGAAAUUACAUUGAUUUAAGA
>mmu-mir-5100_5p
UGGGAGGGAGGACUUGGGAA
>mmu-mir-5101_5p
UUUUCUUAGUAUCAGUUACA
>mmu-mir-5103_5p
UUUGGGGACCCUAGGAUCUGGG
>mmu-mir-5104_5p
UGAGGCAUCUCUCUAGCUCCAGA
>mmu-mir-5106_5p
CAACAACAACAGCAACAACCCG
>mmu-mir-5108_5p
UCCACUGUUCUUACCAUUCCCU
>mmu-mir-5109_5p
CCGUGCCUGGGCUGACACCUAG
>mmu-mir-5118_5p
GAGGCAGAGGUUGGCUGAUCU
>mmu-mir-5119_5p
CAGGGCUGGCCUAUGGGACAGA
>mmu-mir-5121_5p
AUGUGGUGACAUGUAGGACAGG
>mmu-mir-466q_5p
UGUGUGUGUGUGUGUGUAUGU
>mmu-mir-5123_5p
CAUAUGCCAUGGUGUGUAGAU
>mmu-mir-5125_5p
UAAGGAGAGCCCCAUGCCUUUG
>mmu-mir-5126_5p
ACGCCUCCUGCAGCUGCGGGAG
>mmu-mir-5127_5p
UGGUGAAAUGUGGUGACAUAU
>mmu-mir-5128_5p
GUCUUUCUAGCUCCUGUUUUAC
>mmu-mir-5131_5p
UCGGAUGCGCGUGUGGCGGAAG
>mmu-mir-5135_5p
ACUCGGAGCCCAGCCACCUAGA
>mmu-mir-5136_5p
GGUGGUUUUCACGUGAGUCUUG
>mmu-mir-6236_5p
UGAAAAUGGAUGGCGCUGGAGCG
>mmu-mir-6237_5p
UGGGACAGGACUCAACACUCA
>mmu-mir-6238_5p
UAUCUGACCUGAGUGAACUAGG
>mmu-mir-6239_5p
CCUGGGUGUUAGCGUUGGAUC
>mmu-mir-133c_5p
AGGCUCAUGAAGACACCAAA
>mmu-mir-6337_5p

UCUAACAUUUCACAGUCCUUUUC
>mmu-mir-6342_5p
UCCGCUAUCAUGCAACUGUGGC
>mmu-mir-6343_5p
UUCAGAUAGCUAUUUAUUGC
>mmu-mir-6346_5p
AGGGCAGAUGACCCAAGUCCAG
>mmu-mir-6348_5p
GGCACAAAAGGAAAGGUGGUU
>mmu-mir-6349_5p
AAUUCUUUAAUCCCUCCCACC
>mmu-mir-6357_5p
AGGCGCGCUUGUCGGUGUAGGC
>mmu-mir-6359_5p
ACUGAUUGUGGCGGAGAGAUGG
>mmu-mir-6361_5p
CAUGGGGCUGAAUACUGUUGGG
>mmu-mir-6363_5p
UCGAGAUAAAAUUCCUCUACGUG
>mmu-mir-6364_5p
UAUAGGGCUGGGGAAUAGCUU
>mmu-mir-6365_5p
UCUAAUAGUAGAAGACCUGGGU
>mmu-mir-6366_5p
UGGCAAGUGGGUCUCCUGGGGA
>mmu-mir-6367_5p
AAGCAAGAACAGGAGUUCAAAGG
>mmu-mir-6368_5p
CUUCACUCUGUUAUUUCCUAGA
>mmu-mir-6370_5p
UUCAUAUGUUGGCUGCUGAUGU
>mmu-mir-6371_5p
AGUAGUUUGUCAUGCUGGAGGU
>mmu-mir-6373_5p
GCUUUCUAAUACUCAUUUUUCA
>mmu-mir-6377_5p
UAGAAGAGAAACCUGAAGUACU
>mmu-mir-6379_5p
UAGCAGUGAAAGCUUUGGGAA
>mmu-mir-6380_5p
GGUUGUAAGCAACUAUGUGGAUA
>mmu-mir-6382_5p
ACCACUUGGUUCUCUGUCACAU
>mmu-mir-130c_5p
UGGAUGUAAAAUGUCCCCUGCA
>mmu-mir-6391_5p
UGGGCAUCCUCCCAGAGUCUAG
>mmu-mir-6394_5p
AGCUGUUGCCUUUCCCAGGGG
>mmu-mir-6396_5p
ACAGGUGACUAGGGGCGAAGG
>mmu-mir-6397_5p
UUCUCACUGAAAGAAUCCUGGA
>mmu-mir-6398_5p
UUUUUGGGGGGGGGCAUAUAUUC
>mmu-mir-6399_5p
UCAGAUACCAUUGUUUGAUCC
>mmu-mir-6400_5p
CGGCUGUGUCUUGCAGGAAGC
>mmu-mir-6401_5p
AAUGAUACCAUAACUGGGCACUG
>mmu-mir-6402_5p
UGGGAAUAUUAUAACUGUUUAG
>mmu-mir-6403_5p
UGUGCGGCCGCUGCCUCUGUCA
>mmu-mir-6404_5p
UACAUCAUGUCCCAUCACUAGA
>mmu-mir-6405_5p
UCCAAGGGCUUCUUUCUGACUU
>mmu-mir-496b_5p
UUGGAAGCAGAUGGCCGAUAAUU
>mmu-mir-6407_5p
UGCCACAUUGCAUUCUGGGGAG
>mmu-mir-6408_5p
AGGGACAUUCUGUUCAAGCUCA
>mmu-mir-6409_5p
CAAUAACAGCUAGCGCGCAUGC
>mmu-mir-6412_5p
UUAGCUUGAUGUGGUACUGCAC
>mmu-mir-6413_5p
AGCCCUUGCCUGUCCUUGCCUAA
>mmu-mir-6414_5p
AGGCUUUUCAGACCCUGCUCUUU
>mmu-mir-6420_5p
UGGGGGUGGGAUGGAGUGGGGA
>mmu-mir-873b_5p
UGUGGGUGUUCCCGGGAACUUG
>mmu-mir-6541_5p
UGAGAGAGUCCUUGCCUGAGCA
>mmu-mir-692-3_5p
AGACUGGCGCGCCCCAGGGAUCU
>mmu-mir-297a-1_3p
UAUUGCAUGUAUAUAUUAUGC
>mmu-mir-297a-2_3p
CAUACCCAUACAAGCAUGCAC
>mmu-mir-451a_3p
UAGUAAUGGUAACGGUUCU
>mmu-mir-484_3p
UUACCUAGGGGGCUGGCGGCGU
>mmu-mir-546_3p
UUGUCUCUUGCUAUCCCUGUGC
>mmu-mir-761_3p
UUUCACUUUGCUGCUCCUCCUG
>mmu-mir-763_3p
UCUGCCUCCCAGCCAGCCAUUA
>mmu-mir-759_3p
UAAAUGUUUGCACUGGCUGUUU
>mmu-mir-680-1_3p
AUCCUCUUGACAGCCUUGGGU
>mmu-mir-680-3_3p
UAGGCAGCAGGUACUCUUCAU
>mmu-mir-681_3p
CAGGGCCUCCAGCGGGACAGUU
>mmu-mir-686_3p
UGGGCACCAUGGCUGGGGGUG
>mmu-mir-687_3p
AGUCUGUCAUUGUAUUCUUG
>mmu-mir-695_3p

UUGGUCCUGGUCACCGGCUCGG
>mmu-mir-703_3p
UUUUUUUUUGUGUGUGGCAGUU
>mmu-mir-705_3p
CAGUCCUCCCUACCUCCCUAAC
>mmu-mir-706_3p
UUUUUUGAGAUGGCUUUUUUUU
>mmu-mir-707_3p
AUGGGCAUGCGGUCAUAGUUGU
>mmu-mir-710_3p
UUCAACUCUUUAGAACUUAGGU
>mmu-mir-714_3p
GGUUGGCGCGGUCGCCCGGCGC
>mmu-mir-717_3p
GAAGCUGCUCUCCGUUCCGAAG
>mmu-mir-721_3p
UCAUUUUUCUUGUUAUUGCCAC
>mmu-mir-882_3p
UGAUUUCUGGGUUUUUCUAAU
>mmu-mir-105_3p
ACGAAUGCUUGAGCAUGUGCUA
>mmu-mir-327_3p
GUCCAACAUCCCUCUUGAUGGC
>mmu-mir-568_3p
CGGUGUGUGUAUUAUACAGGU
>mmu-mir-449b_3p
CAGCCACAGCUACCCUGCCACU
>mmu-mir-466f-4_3p
UGCAUGCAUGUGUGUGCAUCUAU
>mmu-mir-466k_3p
GUAUGUGAAUAUAUGUGUAUA
>mmu-mir-1195_3p
CUAGGGCAGCCAGGAACACACA
>mmu-mir-1903_3p
CUCCUGGAAGAGGAACAAGUGU
>mmu-mir-1899_3p
UUUCAGAUUCUGCUCAUUCCGU
>mmu-mir-1904_3p
UCUCUUCAGGUAGAUUAAACAU
>mmu-mir-1898_3p
UAUAAUCUGUUUACUUUGACCUA
>mmu-mir-1935_3p
CCAGCCUGGUCUACAGAGUGA
>mmu-mir-1938_3p
CUGAACUGCAGUUCCCAUCAUG
>mmu-mir-1940_3p
CCUUCCAUUGGUUAAGACCUCC
>mmu-mir-1946a_3p
AAGGCAUGCGCCACCACUCUCG
>mmu-mir-1950_3p
UCAUCUCCUUAAAUGCAGAAA
>mmu-mir-1952_3p
UCCAAGGUGUGGAUGACAAA
>mmu-mir-1953_3p
GAAGGCUGUGAGGUUCCCCUCU
>mmu-mir-1954_3p
GACGGGGUUUCUCUGUGUAGCCC
>mmu-mir-669n_3p
UGUGCAUCCCACAGCACAUGU
>mmu-mir-1956_3p
UUCCCUGGCUGGCACCUGGACGU
>mmu-mir-1957a_3p
CAUGUGCAAGGCCCUGAGUUU
>mmu-mir-1960_3p
GCUCUCUUCUGCAGCACUGACU
>mmu-mir-1963_3p
AGGACUCGGCCCUUGUCCCGCA
>mmu-mir-1969_3p
AUCCAUGCUGCCUCCAUUUCUG
>mmu-mir-1970_3p
CAGUCAGGCCUAGUGGCACUCA
>mmu-mir-1971_3p
UCUGUUUUUCAGUCUAUCU
>mmu-mir-2183_3p
AUGGGGUUCUAAGAAUCUGCA
>mmu-mir-767_3p
UCUGCUCAUACCCUAUGGUUCCU
>mmu-mir-3470a_3p
AUCAGCCUGCCUCUGCCUCCU
>mmu-mir-3470b_3p
CCUGCCUCUGCCUCCCGAGUGC
>mmu-mir-3471-1_3p
UGCUCUGUCCAGUUUCUUUUU
>mmu-mir-3471-2_3p
CCCCUUGUGGGUGGGACAAUCU
>mmu-mir-1186b_3p
UGGUGCCUGACUGUGAUCCCAA
>mmu-mir-3474_3p
AUCUGCGUCUUGUUCCAGGUUC
>mmu-mir-3963_3p
UUACAGGUUUUAGGUGGAAUAU
>mmu-mir-3966_3p
UGUUCAUCAUGCUUGCUGCAGA
>mmu-mir-3969_3p
UUAGAUUAGCUACUUACAGGGC
>mmu-mir-28b_3p
CCAGAAUGUGUGAGGCAUCUU
>mmu-mir-3970_3p
AGCAGAAACCAGCAUCACCCUU
>mmu-mir-5046_3p
UCCGAUCCGGGAGCCUGGU
>mmu-mir-5098_3p
CAGGGUUUCUCUGUGUAGCCC
>mmu-mir-5105_3p
CCCAGCCCGUGGACGGUGUGA
>mmu-mir-3473c_3p
UUAUAGGAGUUGGGGAGAUG
>mmu-mir-5110_3p
UUCUAGCCACUUGCCCUGAGU
>mmu-mir-5112_3p
UGCUUCAUCCCCAGCUACA
>mmu-mir-5113_3p
AGGGUCACUCCCUCCUCUCUGC
>mmu-mir-5114_3p
UUGCAGUUCCUGUUCCAGAAG
>mmu-mir-5115_3p
CCUCAGCUGCGGUGGGUGUCA
>mmu-mir-5116_3p

AGGCAGGGUCUUCAUAUCGAGA
>mmu-mir-5120_3p
UCAGUGGCAGCAGCCCCUUCAG
>mmu-mir-5122_3p
AAGGAGCUGCCGCUGGGCCG
>mmu-mir-3473d_3p
AGGGCUGGAGAGGUGGCUCAGU
>mmu-mir-5124a_3p
UCUUGCAGAGGACCCAAGUUCA
>mmu-mir-5130_3p
CUCGCACCGCGCGGCUCUCAG
>mmu-mir-5133_3p
UGUCUCUGCCGCUCGCUUCAG
>mmu-mir-344i_3p
GGCUCUAGCCAGGGUCUGACUAC
>mmu-mir-5709_3p
AAAGUCUUAAGGGUGUGUAUUG
>mmu-mir-5710_3p
CUGGCUCUUUGUCCCUCGGCA
>mmu-mir-6240_3p
UGUGAUUUCUGCCCAGUGCUC
>mmu-mir-6241_3p
UGAGGGAAUUUCCAGGUGGCCA
>mmu-mir-6243_3p
UCGGUUGGCCCCGGAUAGCCGG
>mmu-mir-6244_3p
CCGAUAGCCAUAUGCUUCCAG
>mmu-mir-195b_3p
AAUUACUGCUCUCCUGUAACU
>mmu-mir-6335_3p
UAGAAUGAGUUACAUGAGGAC
>mmu-mir-6336_3p
UCUUUUAUAGGUCUGAUGGAAA
>mmu-mir-6338_3p
UAGAAGGGAGAAUGUAUGA
>mmu-mir-6339_3p
AAGCUUGGGAAGCUGGUCUCCU
>mmu-mir-6340_3p
AGUUUGAAGCCCUUGCCGCCAG
>mmu-mir-6341_3p
UAGUAUAACACUGAGGGUCAAC
>mmu-mir-6344_3p
CAGAGUGGAGAUGGGAGAACAG
>mmu-mir-6345_3p
UACUAGGUUCUCCAUGGACA
>mmu-mir-6347_3p
UGCAGUCUUUCACAAGCUCAC
>mmu-mir-6350_3p
AUAUUUUCAUGCCUUGAGCAC
>mmu-mir-6351_3p
AGAGAUCCUCUGGGCAUGCUC
>mmu-mir-6352_3p
UGUGGAAUCUUGCUGUUCCUUC
>mmu-mir-6353_3p
AGAGCCAGGAGUGUGUGUCUGG
>mmu-mir-6354_3p
AGAGACUGCAACCAGGAAGUCU
>mmu-mir-6355_3p
UUUUAUUAUGAUGCUGAUUGU
>mmu-mir-6356_3p
CAGCAGGGCACUGUGCAGGAA
>mmu-mir-6358_3p
UGAUGGUUUGUAUAUCCUUGGA
>mmu-mir-6360_3p
CAUCUGGGGGUCAUAGGUCAAC
>mmu-mir-145b_3p
UCUGGCUUGAGAAAUUGGUGU
>mmu-mir-6362_3p
AGCAGGGUGUGGAGAGUCCUU
>mmu-mir-6369_3p
UAUGUAGCAGAGGAUAGCCUAG
>mmu-mir-6372_3p
UGCUGGUUUUCUAUGUCCAUC
>mmu-mir-6374_3p
UAAAAACUAUUAUGGUUUUCU
>mmu-mir-6375_3p
UGAUACCAGCUACCACAGUGU
>mmu-mir-6376_3p
CAUCACAGUUUCUAAUGCUCAGC
>mmu-mir-21b_3p
AGAAAAUCCUUCUGUACUAUCU
>mmu-let-7j_3p
CCCUUGCUCAGAUUAAAAGCCUGG
>mmu-mir-6378_3p
CUGAUUUUCUCUGAAAUAUGG
>mmu-mir-6381_3p
UCCUCUUCCCAGCUAGCUUGU
>mmu-mir-6383_3p
UCAGUGGGCAAUACUAUGCUA
>mmu-mir-6384_3p
GUGAGCCACGUGGGGAGAUGGA
>mmu-mir-6385_3p
UUUCCUCAUGUAUCUGGGCCC
>mmu-mir-6386_3p
GGGAUUCUCACUUGCUGAUGCA
>mmu-mir-6387_3p
CCCACAAUGCCGGAGGCUCCAU
>mmu-mir-6388_3p
AGCAGACUGUGUUCUCUGUU
>mmu-mir-5124b_3p
GCAUACUCAUUGGGUAGUUCUA
>mmu-mir-6389_3p
UGUAUUUGUGUUCAACAGCUUU
>mmu-mir-378c_3p
CAACAUGUAGUCUAUCUGAUCU
>mmu-mir-6390_3p
UUUGAUGCUUCGAAAUCUUUU
>mmu-mir-6393_3p
CAAGUGUCACUGUUGGCCUGG
>mmu-mir-1957b_3p
UGCAAGGCGCUGAGUUCCAGCC
>mmu-mir-6395_3p
CAGACAUGAGUCUUCUGUCAGC
>mmu-mir-21c_3p
CAGCUCUGUUCAGCUAUUCUCA
>mmu-mir-6406_3p
CAUGUGGCUGGCACUCCAGGAA
>mmu-mir-6410_3p

GCCGAGGAGUUCCAUGGCCCUG
>mmu-mir-6411_3p
Not_Predicted
>mmu-mir-378d_3p
CUGUGAUUUUUAGGGUGUCAGU
>mmu-mir-6415_3p
CAUUACACUUUGAAGAGUCUCC
>mmu-mir-6417_3p
UUGUGUGCUGCUGUGAUUGUCC
>mmu-mir-6419_3p
UGUUGCACGUGCUGCUGAGUCU
>mmu-mir-451b_3p
GGUUUCCUCGCCAUUCCCAAG
>mmu-mir-30f_3p
GCUUCCAGUCAAGGAUGUUUAC
>mmu-mir-3473e_3p
GAUGCUUUCUCAGAGGACCCAA

Karathanasis Nestoras

CURRICULUM VITAE

Born in Athens on 18/09/1983
41 Iktinou street Artemida, Greece

E-MAIL nk3932@gmail.gr

EDUCATION

1. **Current position:** PhD thesis on miRNAs and cancer. An experimental and Bioinformatics approach.
2. **Master's degree:** in Molecular Biology-Biomedicine. Biology department, University of Crete.(8.83/ 10)
3. **Bachelor:** Department of Molecular Biology and Genetics, Democritus university of Thrace (Alexandroupoli, Greece). (7.81/ 10)
4. Highschool (18,1/20)

PREVIOUS EXPERIENCE/ SCHOLARSHIPS

1. **PhD thesis:** Currently, I am completing my PhD thesis in Dr. Poirazi's lab at the Institute of Molecular biology and Biotechnology (IMBB), which is co-funded by a scholarship from the research program Herakleitos II.
2. **Master thesis:** I performed my master thesis in Dr.Poirazi's lab, at the Institute of Molecular biology and Biotechnology (IMBB), and as a scholar in Dr. Kelsey's Martin lab in the department of Psychiatry and Biological Chemistry at the University of California, Los Angeles.
3. **Bachelor thesis:** I performed my thesis project, as a scholar, in the Center for Research on Reproduction and Women's Health at the University of Pennsylvania (Dr. Coukos's laboratory).
4. Practical training in the gene expression lab of Dr. Kretsobali (IMBB).
5. Practical training in the immunology lab of Dr. Athanassaki (Biology Department, University of Crete).

PUBLICATIONS

Research papers

1. **Karathanasis N**, Tsamardinos I, Poirazi P MiRduplexSVM: a high-performing miRNA-duplex prediction methodology. Submitted to NAR
2. **Karathanasis N**, Tsamardinos I, Poirazi P Don't use a cannon to kill the … miRNA mosquito. Submitted to Bioinformatics

3. **Karathanasis N,** Oulas A, Louloupi A, Iliopoulos I, Kalantidis K, Poirazi P. A new microRNA target prediction tool identifies a novel interaction of a putative miRNA with CCND2.: RNA Biology, 2012 Sep;9(9):1196-207

Review papers – Book chapters

4. Oulas A, **Karathanasis N**, Louloupi A, Poirazi P. Finding Cancer-Associated miRNAs: Methods and Tools. Mol Biotechnol. 2011 Sep;49(1):97-107.
5. Oulas A, **Karathanasis N**, Poirazi P. Computational identification of miRNAs involved in cancer. Methods Mol Biol. 2011;676:23-41.

PUBLICATIONS

Conference paper

6. **Karathanasis N**, Tsamardinos I. Poirazi P. A bioinformatics approach for investigating the determinants of Drosha processing. 2013 IEEE 13th International Conference on Bioinformatics & Bioengineering (BIBE)
7. **Karathanasis N**, Angelos A. Tsamardinos I. Poirazi P. SVM-based miRNA: MiRNA∗ duplex prediction. 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)

PRESENTATIONS / CONFERENCES

1. **Nestoras Karathanasis**, Angelos Armen, Ioannis Tsamardinos, Panayiota Poirazi. miRNA:miRNA* Duplex Prediction using a SVM approach. 63RD Congress of the Hellenic Society of Biochemistry and Molecular Biology. 2012
2. **Nestoras Karathanasis**, Angelos Armen, Ioannis Tsamardinos, Panayiota Poirazi. DuplexSVM: a miRNA-duplex prediction tool. 7th Conference of the Hellenic Society for Computational Biology & Bioinformatics. 2012
3. **Nestoras Karathanasis,** Anastasis Oulas, Annita Louloupi, Ioannis Iliopoulos, Kriton Kalantidis and Panayiota Poirazi. A new microRNA target prediction tool identifies a novel interaction of a putative miRNA with CCND2. 7th microsymposium of small RNAs. 2012
4. Anastasis Oulas, **Nestoras Karathanasis**, Ioannis Iliopoulos and Panayiota Poirazi. Prediction of miRNA gene targets – a combined computational and experimental approach, ISMB 2011

5. **Karathanasis Nestoras**, Kelsey C. Martin, Panayiota Poirazi "mRNA-miRNA predicted interactions related with synapticplastisity" The Seventh Annual Southern California Learning & Memory Symposium Neuroscience Research Building Auditorium, UCLA June 4, 2008

LANGUAGES Greek (native), English (fluent)

MOLECULAR BIOLOGY TECHNIQUES

- Northern
- Isolation of DNA/RNA,
- Polymerase chain reaction (PCR),
- Real-Time PCR
- Colony PCR
- Reverse Transcription
- Primer Extention analyses
- Electrophoretic analysis of proteins and nucleic acids,
- Molecular cloning techniques (genetic engineering, transformation, minipreps, bacterial cultures, etc),
- Isolation and characterization of protein based on enzymatic activities
- Light Microscopy
- Expression and purification of recombinant proteins,
- Immunological assays (ELISA, immunostaining, western blot, immunoprecipitation)
- Luciferase assay
- Cell cultures
- Radioactivity Reactions

COMPUTER SKILLS

1. Machine Learning
2. Java computer programming language
3. Algorithms in Bioinformatics
4. Matlab
5. Molecular graphics (Rasmol)
6. Database searching (NCBI)
7. Basic sequence alignment/analysis (BLAST)
8. MS Office software

TEAM WORK PROJECT

During my Bachelor I participated in the following team work projects, involving the presentation of advanced topics of molecular biology and research articles:

- Gene therapy for H.I.V.
- Function and structure of lambda repressor
- Chemical analysis of oil
- Soap
- Are we determined by our genes?
- Chromosome engineering in mice
- Dlx proteins position the neural plate border and determine adjacent cell fates

PERSONALITY

Cooperative, well-organised, diligent, persistent, flexible and adaptive, fast-learner

CAREER OBJECTIVES

- To pursue a challenging research project leading to a PhD degree.
- To contribute to the understanding of incurable diseases through the use of molecular biology technology.

OTHER INTEREST

Sports, cinema, music