Silencing the Host – The Role of Intronic microRNAs

by

Ludwig Christian Giuseppe Hinske

Medical Doctor
Ludwig-Maximilians-Universität München, Germany, 2007

SUBMITTED TO THE DEPARTMENT OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN BIOMEDICAL INFORMATICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2009

Signature of Author:

Department of Health Sciences and Technology
December 1st, 2008

Certified by:

Lucila Ohno-Machado, MD, Ph.D.
Associate Professor of Radiology and Health Sciences and Technology
Thesis Supervisor

Accepted by:

Lee Gehrke, Ph.D.
Hermann von Helmholtz Professor of Health Sciences and Technology
Interim Director, Harvard-MIT Division of Health Sciences and Technology

Silencing the Host – The Role of Intronic microRNAs

by

Ludwig Christian Giuseppe Hinske

Submitted to the Department of Health Sciences and Technology
on January 15, 2009 in Partial Fulfillment of the
Requirements for the Degree of Master of Science in
Biomedical Informatics

ABSTRACT

Fifteen years ago lin-4 was reported to be the first endogenous small non-coding, but interfering RNA structure involved in developmental timing in C. elegans. First thought not, or only rarely, to occur in mammals, microRNAs are now among the major players in up-to-date genomic research. The mature molecules are ~22 nucleotides in length and, by targeting predominantly the 3' UTR of mRNAs, lead to translational repression or degradation of the target message, hence controlling important cellular mechanisms, including division, differentiation and death. This key role makes them excellent targets for cancer research. In fact they have been shown to have a major impact on cancer development in many cases. However, miRNAs are not a homogeneous class and can be subclassified into intragenic and intergenic, depending on their genomic position. Whereas intergenic miRNAs are expected to be independent transcriptional units, intragenic miRNAs are commonly believed to be regulated through their host gene. Despite of the growing knowledge on how miRNAs integrate into cellular regulatory networks, our current knowledge about the specific role of intragenic miRNAs is rather limited. In this work we integrated current miRNA knowledge bases, ranging from miRNA sequence and genomic localization information to target prediction, with biochemical pathway information and publicly available expression data to investigate functional properties of intragenic miRNAs and their relationship to their host genes. To the best of our knowledge, we are the first to show in a large-scale analysis that intragenic miRNAs seem to act as negative feedback regulators on multiple levels. We furthermore investigated the impact of this model on the potential role of intronic miRNAs in cancer pathogenesis.

Thesis Supervisor: Lucila Ohno-Machado
Title: Associate Professor of Radiology and Health Sciences and Technology

ACKNOWLEDGEMENTS

This thesis is dedicated to my family, including my wife and best friend Patricia, my parents Annette and Ludwig, my brother Nicolas and my sister Stephanie. It is their love and endless support that made my stay and work possible.

It is impossible to overstate my gratitude to my advisor, Professor Lucila Ohno-Machado, who with her infectious enthusiasm, warm personality, and endless patience guided me, academically as well as personally.

I will miss the stimulating discussions through day and night with Doctor Pedro Galante, who was a friend and a teacher to me, helping me to understand crucial principles of bioinformatics.

The warm and welcoming climate in my laboratory, the Decision Systems Group, was a wonderful environment and I can hardly imagine greater colleagues to work with. I owe special gratitude to Erik Pitzer and Jihoon Kim, who would always spend as much time as necessary to explain algorithms or statistics to me.

Finally, I owe my deepest gratitude to my friends. It was their camaraderie and constant support that made Boston a home for me during my time here. I would like to explicitly thank Leo Celi and Guido Davidzon, Richard Lu, Ben Geissler, Christopher Tsai, David Singerman, Michael Parker, Pankaj Sarin, Christine Hsieh, Tanguy Chau, Minsue Suh, Alberto Ortega and Bronwyn Wyatt.

CONTENTS

FIGURES

TABLES

1 Introduction

1.1 MicroRNA – Overview

Critical cell functions are carried out by proteins. The information on how to assemble proteins from their basic chemical structure, namely amino acids, is stored in genes, defined regions within the DNA, which can be collectively referred to as the 'genome'. Similarly, the set of all proteins in an organism is called the 'proteome'. To produce a protein, the information on the DNA is read by an RNA polymerase that will transcribe a temporary message, the messenger RNA (mRNA), which in turn gets translated into a protein.

In 1993, Lee, Feinbaum and Ambros found that, in *C. elegans*, the gene lin-4 did not encode a protein, but rather a small RNA that would interfere with protein levels of lin-14 [1]. Based on the involvement in heterochronic pathways, these molecules were first dubbed small temporal RNAs (stRNAs) [2]. Almost seven years later, Pasquinelli and coworkers were able to identify the small, non-coding RNA *let-7* in multiple species, including *Homo sapiens*, leading to speculation that probably more molecules of a similar kind would be detected [3-5]. This would come true within a year, when Lagos-Quintana et al., Lau et al. and Lee and Ambros successfully cloned several new genes with similar properties. In contrast to lin-4 and let-7, however, many of these genes could not be linked to temporal development, so the name 'microRNA' (miRNA) was established [6]. The major properties of miRNAs are that they are processed from a precursor that contains a hairpin structure, that their active form is a single-stranded RNA molecule of ~22 nucleotides in length, and that they seem to primarily bind to the 3'-untranslated region (UTR) of certain mRNAs, modulating protein levels.

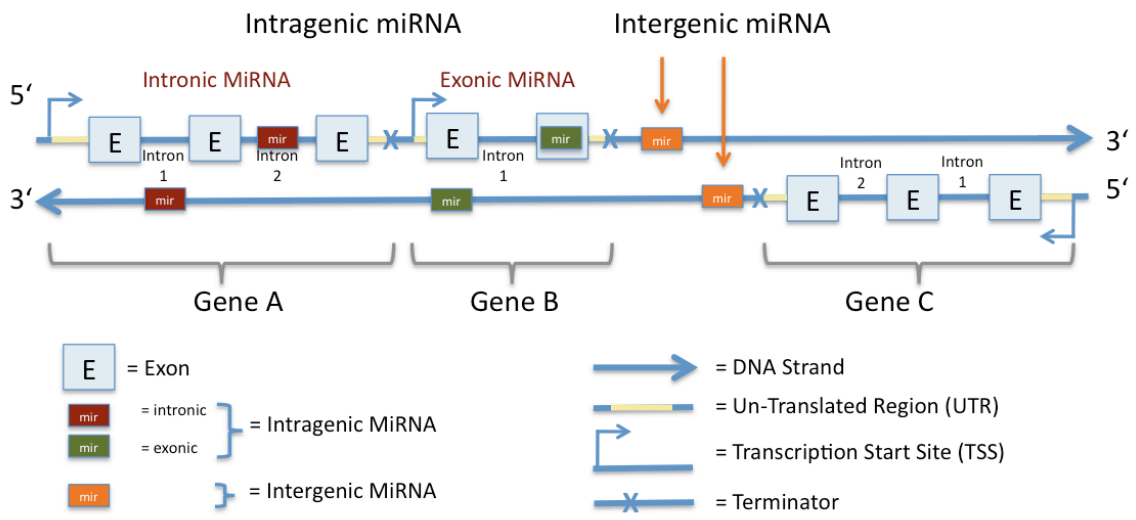

Figure 1 - Classes of miRNA

Depending on the genomic position, miRNAs can be classified into intragenic and intergenic. Intragenic miRNAs can further be subdivided into intronic and exonic. Whereas most intragenic miRNA genes are on the same strand as their host genes, a few reside on the opposite strand. In this work, only protein coding genes were considered as hosts for miRNAs.

The miRNA class of molecules is not homogeneous, however. Whereas about half of human miRNAs are intergenic, i.e. found in distant locations from currently annotated genes, the other half of currently known miRNA genes are intragenic, i.e. located within protein coding genes. Intragenic miRNAs can be subdivided into intronic and exonic, as shown in Figure 1. Most miRNA genes are on the same strand as their host genes, suggesting common regulation [2, 7]. Some intergenic miRNAs are clustered and believed to be transcribed as a polycistron [7].

1.2 miRNA Biogenesis

The process of miRNA processing and cleavage is largely understood. Most miRNAs are transcribed by the polymerase *Pol II* with typical features such as a polyadenylated tail and a 5'-cap structure [8, 9]. Few are transcribed by *Pol III* (mainly those that reside in Alu repeats), including some intronic miRNAs [10]. The resulting transcriptional product is called the primary miRNA (pri-miRNA) and can vary greatly in length, up to tens of thousands of nucleotides. The pri-miRNA forms a hairpin loop structure that undergoes further processing in the so-called "microprocessor", a protein complex including the

RNA binding enzyme *DGCR8* and the RNAse III *Drosha* [8, 11-13]. *Drosha* cuts the double-stranded end, leaving a ~70 nucleotide long hairpin precursor miRNA (pre-miRNA) [14]. As is typical for RNAse III cleavage, the pre-miRNA contains a 2 nucleotide 5'-end overhang that is recognized by Exportin 5, which is necessary for transport into the cytoplasm [15-17]. In contrast to intronic small nucleolar RNA that is extracted after the splicing process of its host mRNA [18], Kim et al. recently showed that an intronic miRNA can be extracted from its intron before splicing occurs and without affecting translation of its host mRNA [19].

In a second processing step, a protein complex including the RNA recognizing protein *TRBP* and another enzyme of the RNAse III family, Dicer, cuts out the hairpin loop structure, leaving the mature miRNA:miRNA* double strand [20-24]. Usually, one of the two strands will be degraded, whereas the other is incorporated into the so-called RNA-induced silencing complex (RISC). Which strand will be the active one depends on the relative and absolute stability of 5'-base pairing [25, 26].

1.3 miRNA Target Interaction

The miRNA incorporated in RISC recognizes its target through Watson-Crick complementarity of its 5'-end to the 3'- UTR of its target, and details of this process have recently been identified [27]. Whereas in plants miRNAs seem to nearly perfectly match the target sequence, this is not true in mammals, where imperfect pairing is predominant and near-perfect complementarity is only required for the "seed-region" of the mature miRNA (nucleotides 2-8). After recognition, RISC 'silences' its mRNA target through either translational repression, degradation, cleavage or storage in so-called P-bodies, ribosome-less, cytoplasmic structures (reviewed in [28]). Figure 2 illustrates the basic mechanism of miRNA target interaction.

So far, the literature suggests that at least four different mechanisms may explain the underlying nature of miRNA-induced translational repression. In 2002, Seggerson et al. [29] observed that miRNAs and their targets were associated with polysomes that seemed to be actively translating target mRNA. Similar results were also found by [30-32], which

led to the proposal that miRNAs might be involved in co-degradation of the evolving polypeptide chain. However, until today the identity of the protease that would be required for such a process remains unknown [28]. Based on a reporter essay, Petersen suggested that translational repression might be promoted through premature polysome dissociation [32].

Whereas the previous studies displayed evidence for post-initiation translational repression, Kiriakidou et al. could show that Argonaute competes with eukaryotic translation initiation factor 4E (*eIF4E*) for mRNA cap structures that play an important role in translation initiation [33]. Another way of repressing translation initiation was proposed by Chendrimada et al., whose results suggested that *AGO2* might recruit *eIF6* and hence prevent association of ribosomal subunits [34].

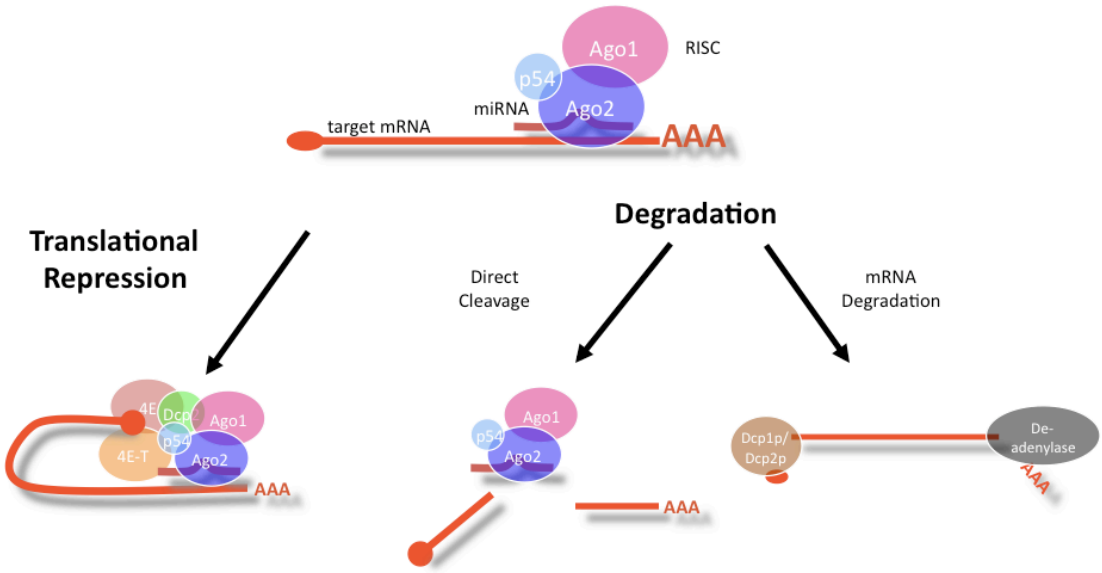

Figure 2 - miRNA Target Interaction

The mature miRNA single-strand molecule is integrated into RISC. The target is detected by complementarity of the miRNA 5' region to the 3'-UTR of its target mRNA, followed by either translational repression or degradation of the mRNA molecule.

Whereas it was originally observed that miRNAs repress the target gene protein levels without affecting mRNA levels [1], others were able to show that animal miRNAs

significantly reduce expression levels of targeted mRNA. This allowed the development of certain target prediction algorithms [35] that were based on systematic miRNA over-expression experiments [36]. Whereas in plant-miRNAs endonucleolytic cleavage by *Argonaute* proteins seems to be the prevailing mechanism of mRNA degradation, in animal cells mRNAs are processed by the general mRNA degradation machinery, including accelerated deadenylation and decapping [37-40]. It is interesting to note that Wu et al. were able to show that mRNA degradation and translational repression can be uncoupled from one another, suggesting independent mechanisms [39].

1.4 miRNAs in Human Disease

Due to their substantial role as regulatory elements, it is not surprising that miRNAs were identified as playing central roles in diverse classes of diseases, including cancer, infections, muscle conditions and neurologic diseases.

In tumors, proto-oncogenic as well as tumor suppressing miRNAs have been reported. For example, Li Ma, Julie Teruya-Feldstein and Robert Weinberg deciphered the mechanism of how *miR-10b* over-expression promotes cell migration and invasion in breast cancer. They found that the transcription factor "Twist" positively regulates *miR-10b*, which in turn inhibits translation of homeobox D10. This leads to increased expression of the pro-metastatic gene RHOC. They also showed that *miR-10b* over-expression correlates with clinical outcome [41]. Similar findings have been reported for other miRNAs and other tumor types, including *miR-21* in colorectal adenocarcinoma [42] and breast cancer [43] or *miR-155* in lymphatic malignancies [44] and pancreatic cancer [45]. Recently, Tavazoie showed that *miR-126* inhibits tumor growth and *miR-335* reduces metastatic spread through targeting of the transcription factor *SOX4* and the extracellular matrix component tenascin C in breast cancer [46]. Likewise, tumor-suppressive miRNAs have been identified in other tumors [47, 48].

Interestingly, viral genomes encode miRNAs to modify their hosts microenvironment, including herpes simplex virus [49], human cytomegalovirus [50], Eppstein-Barr virus

[51, 52], and HIV [53]. Recent evidence accumulates that in fact cell-encoded miRNAs may be able to regulate viral mRNA [54-58].

It is known that certain miRNAs are preferentially expressed in certain tissues. Tissue specificity is defined as a greater 20-fold increase in expression levels of a certain miRNA as compared to other tissues [59]. Heart and skeletal muscle, brain and pancreas tissue contain the largest number of tissue-specific miRNAs known to date [59]. Recently, Eisenberg et al. reported multiple differentially expressed miRNAs in primary muscle disorders [60]. Similarly, Carè and coworkers discovered that inhibition of *miR-1* and *miR-133* may induce cardiac hypertrophy [61] and evidence accumulates for a significant role of miRNAs in myocardial remodeling [62].

Neurologic disorders comprise another broad class of human diseases displaying pathogenetic association with miRNAs (reviewed in [63]). For example, altered miRNA expression levels were found in patients with schizophrenia [64], Alzheimer's disease [65] and Parkinson's disease [66]. However, so far causal links remain to be identified [63].

1.5 Experimental Methods

1.5.1 Microarrays

Microarrays have been developed in the early 1990s and ever since became more and more popular in the research community. Microarrays can be classified either as oligonucleotide [67] or cDNA arrays [68], which mainly refers to the manufacturing technique, or as single color versus dual color arrays. Two color arrays are designed such that two samples can be hybridized to the same platform (e.g., a tumorous tissue sample and a normal tissue reference sample). Both samples are labeled in different colors, usually cyanine 3 (Cy3, "green channel") and cyanine 5 (Cy5, "red channel"). Single color arrays are hybridized to a single sample. Even though twice as many arrays are needed compared to two color systems, raw measurements allow better comparison across studies. miRNA platforms are slightly different. These are usually self-manufactured academic cDNA oligo arrays, they are also far less dense, as the number of

known miRNA genes in humans approximates 700, opposed to the human genome, consisting roughly of 20,000 genes.

A crucial and yet limiting step in microarray analysis is processing of the raw data, before expression values can be compared. The whole process can be subdivided into background correction, normalization, and summarization. Most platforms do not only provide information on the probe intensities, but also supply a background intensity measurement to help eradicate systematic background measurement errors. Different background correction methods have been proposed [69], the most intuitive being simple subtraction.

Normalization is needed when more than one array is involved in the analysis. It is obvious that if an experiment is repeated on two microarray platforms, the measured expression levels will be similar, but not exactly the same. Introduced systematic biases, due to many reasons including physical properties of the platform, small differences in the preparation of the samples, or chemical behavior of the used fluorescence, add to random variation in gene expression levels, limiting comparability of different arrays. However, if samples are from two different tissues, e.g. cancer versus non-cancer, certain variation in some genes is of great interest and often times the reason for carrying out the experiment. This is referred to as obscuring variation versus interesting variation [70]. The goal of normalization is to reduce as much obscuring variation as necessary while maintaining as much interesting variation as possible. There are many different methods available, both for single color as well as dual color arrays.

A general distinction must be made between complete data methods, i.e. methods using all available arrays for the normalization process, such as quantile normalization and cyclic locally weighted regression and smoothing scatterplots (loess) [70], and methods using baseline arrays, such as scaling and non-linear methods [71].

Quantile normalization [72] is based on the idea that a linear relationship in a quantile-quantile plot of two arrays means that both arrays will have the same distribution of

values. This implies that if one fits the data of multiple arrays to a straight diagonal (in the quantile-quantile plot), the same distribution for every chip will be enforced.

The M versus A plot, where M is the difference in log expression values and A is the average of the log expression values, is the basis of cyclic loess normalization [70]. The plot should be scattered around the horizontal axis. Using loess regression, a normalization curve is fitted to the plot and intensities corrected accordingly. However, this methodology is somewhat computationally intensive, as it requires pairwise iteration through all arrays until the applied changes fall below a certain threshold.
A different set of methods uses a baseline array, which for example is chosen as being the median of all median intensities. Intensities will be corrected by a factor that is the mean intensity of the baseline array over its own mean intensity [73-76]. Whereas scaling methods can be seen as linear interpolation with offset zero, non-linear methods propose an extension to this idea [71].

1.6 miRNA Bioinformatics

1.6.1 miRNA Target Prediction

Until today, there is still no high-throughput method available to identify and validate miRNA targets. Therefore, diverse computational methods have been developed to predict miRNA target interactions. Six commonly used methods are briefly reviewed here: TargetScan, miRanda, PITA, RNA22, MirTarget2, PicTar, as well as TarBase, a database containing experimentally validated targets.

1.6.1.1 TargetScan

In 2003, Lewis et al. [77] presented an algorithm called TargetScan that used secondary RNA structure and cross-species conservation of 3'-UTR motifs as key components to predict miRNA targets. In brief, potential targets are identified by perfect Watson-Crick complementarity of the 5'- seed region of the miRNA to the 3' UTRs of potential target mRNAs. In a second step, the regions in both directions around the seed are aligned using the RNAfold program [78] and a total free binding energy is calculated with the RNAeval algorithm [78]. The free energy is converted to a z-score and predictions in

each organism are ranked. The algorithm takes three parameters: one that defines the relation between the binding energy and the z-score, a z-score cut-off, and a ranking cut-off value. Additionally, it takes cross-species conservation into account. The authors showed that the signal:noise ratio increases from 2:1 (required conservation in human and mouse) to 4.6:1 (required conservation in human, mouse, rat and pufferfish). However, this comes at the cost of significantly fewer predictions. The estimated false positive rate ranges from 22% to 31%, depending on the species and the parameter settings. It is remarkable that even though "TargetScan" is among the earliest published algorithms, it has maintained its role as a gold standard in many experiments.

1.6.1.2 miRanda

The first version of miRanda was developed in 2003 as one of the first miRNA target prediction algorithms by Enright et al. [79]. John et al. [80] adapted the algorithm to predict targets for human miRNAs in the following year. miRanda uses the same basic principles as TargetScan, however the score calculations and parameters are slightly different. Its estimated false positive rate ranges from 24% to 39%, depending on the setting, the number of predicted target sites for a given mRNA 3'-UTR, and the free binding energy score.

1.6.1.3 RNA22

RNA22 is conceptually very different from the algorithms described above. Miranda et al. [81] use the TEIRESIAS variable length motif finding algorithm [82] to derive a list of mature miRNA patterns. Statistical significance of each individual motif is assessed by training a second-order Markov chain. The key idea is that, through the guilt-by-association approach [83], a degree of membership can be calculated for any given putative target site complementary to the motif. Any region that receives more than 30 hits is considered a potential target. The authors use different ways to estimate the false positive rate, which is believed to be between 19% and 26%. Sensitivity estimates range from 36% to 95%, depending on the training dataset. The strength of this approach is that first a sequence in the genome is identified as a potential miRNA target binding site. Theoretically, this enables target identification for miRNAs not yet even known. In an optional second step, the miRNA with the highest degree of membership is selected.

1.6.1.4 PITA

PITA is a relatively new algorithm, published in 2007 by Kertesz and coworkers [84] and mainly based on secondary RNA structure. They could show that site accessibility to the mRNA target site, defined as the energetic cost of resolving intra-mRNA interactions, is as important as seed pairing. In a reporter gene essay, the purely thermodynamic score, which is a combination of the gain in energy by the miRNA binding to the target site and the cost of unpairing the target site's nucleotides, had a high correlation with measured translational repression. They also found that taking into consideration the cost of unpairing 3 nucleotides upstream and 15 nucleotides downstream of the miRNA target site, further significantly improves this correlation. Hence, their algorithm first identifies potential matches by aligning the seed region to the 3' UTR of potential mRNA targets. It then calculates and combines thermodynamic scores for each putative binding site of the miRNA to derive a unique score for a miRNA target interaction. While this method may perform slightly better than PicTar and miRanda, a great advantage is that it does not require cross-species conservation scores or other parameters.

1.6.1.5 MirTarget2

MirTarget2 [35, 85] uses a machine learning approach to target prediction. The key to this method is an experiment by Linsley et al. [36], who systematically studied the change in mRNA expression levels after over-expression of different miRNAs. Wang et al. used this to extract 131 heterogeneous features in the miRNA/target mRNA sequences that correlate with reduced mRNA expression. They then trained a non-linear support vector machine (SVM) on 454 positive samples (down-regulated genes) and 1017 negative samples (unaffected genes). The resulting classifier achieved an Area Under the Receiver Operator Characteristic (ROC) Curve (AUC) of 0.79 in 10-fold cross validation. In transfection experiments, MirTarget2's predictive performance appears to be roughly comparable to TargetScan. The strength of this idea is the utilitzation of biological observation. However, one must keep in mind that observed down-regulation of mRNA could be due to indirect effects, such as downregulation of an enhancing transcription factor.

1.6.1.6 PicTar

PicTar was published in 2005 by Krek and colleagues [86]. Even though it also makes use of successfully employed principles like free binding energy and cross species conservation, the authors increase specificity by reasoning that, similar to transcriptional regulation, co-expressed miRNAs are more likely to target the same mRNAs. Therefore, they use a validated, probabilistic algorithm that has been successfully applied to transcription factor binding site identification [87, 88]. According to the authors, adding probabilistic knowledge about co-expression significantly increases specificity. There exist two different versions of PicTar, the major difference being the number of species for which conservation is required (PicTar 4 requires conservation in human, dog, mouse and rat; PicTar 5 requires also conservation in chicken).

1.6.1.7 TarBase

While the previous methods described are algorithms for computational prediction of targets, TarBase is a database housing manually collected validated miRNA target interactions from different organisms, including human, mouse, fruitfly, worm, and zebrafish [89]. Notably, negative findings are reported as well. Each entry contains the miRNA and target mRNA name associated with the target site, the type of experiment performed, information about whether translational repression or degradation of the transcript was observed, and a reference to the original publication. A drawback is that, due to complex maintenance and current lack of large-scale target validation methods, there are few entries, especially for newly discovered miRNAs.

1.6.1.8 Summary – Target Prediction Algorithms

The different target prediction methods presented above are quite diverse not only in their underlying algorithms, but also in the number of miRNAs predictions are available for and number of genes predicted to be targets. Table 1 gives an overview of the discussed target prediction methods and their main properties.

Table 1 - Overview Target Prediction Algorithms

Target Prediction Algorithm	Total Number of Predictions	% of Known miRNAs	Predicted Targets (% of Known Genes)	Algorithm
TargetScan[77]	1,096,412	67.5%	90.8%	Free binding energy; conservation
miRanda[80]	948,851	97.4%	75.6%	Free binding energy; conservation
RNA22[81]	247,569	46.1%	63.4%	TEIRESIAS motif detection
PITA[84]	4,315,726	97.4%	88.9%	Free binding energy; binding site accessibility
MirTarget2[35, 85]	184,619	74.9%	73.6%	Support Vector Machine classifier
PicTar 5-way[86]	23,089	22.1%	13.6%	Free binding energy; conservation; Co-expression
TarBase[89]**	939	11.6%	2.2%	Experimental validation

* This table is based on our own database, i.e. we considered only predictions where we could match the miRNA symbol to miRBase as well as the target to a gene symbol from NCBI or RefSeq identifier.
** TarBase is strictly speaking not a target prediction algorithm, but a knowledge-base containing information about biologically validated miRNA-mRNA target interactions.

2 Motivation

Until only a few years ago it was commonly believed that intronic DNA regions functioned as spacers and contained little or no meaningful information. However, several authors were able to point out the significance of intronic regulatory elements and their impact on gene expression [90-92]. Some of these correspond to the greater class of miRNAs. miRNAs are single-stranded, ~22 nucleotides long non-coding RNA molecules that, after being processed from a larger hairpin precursor, recognize target mRNA primarily by complementary to its 3'-UTR. Subsequently, the targeted message is predominantly subject to either translational repression or degradation [28], making miRNAs very effective regulatory elements.

Intragenic miRNAs play a unique role within the family of small, non-coding RNAs. Whereas intergenic miRNAs contain their own regulatory elements, including a promoter region and a termination sequence [93, 94], intronic miRNAs are believed to be co-transcribed with their host genes [7]. In this context, Baskerville and colleagues were able to show that expression levels of intronic miRNAs and their hosts were highly correlated in cell line experiments [7, 91], supporting the idea of co-regulation through co-transcription. However, other authors found that, in cancer samples, only a limited number of miRNAs correlated their expression patterns with those of their corresponding host genes [95, 96]. These conflicting findings have been attributed to an altered post-transcriptional regulation of miRNAs in cancer samples [97, 98]. However, the consequences of diverging expression levels of host and intragenic miRNA have not yet been elucidated.

In a recent experiment, Barik [99] has shown that the intronic miRNA *hsa-miR-338* targets a class of mRNAs that are functionally antagonistic to its host, *AATK*. Whereas this experiment suggests functional synergy, it has also been hypothesized that intronic miRNAs could act as negative feedback regulators. Megraw et al. [100] found that, in *Arabidopsis thaliana*, the intergenic miRNAs *ath-miR-160* and *ath-miR-167* may be regulated by *auxin* response factors. Also, other authors have shown that corresponding mRNAs are

targeted by *ath-miR-160* and *ath-miR-167* [89], which therefore down-regulate their own expression. This concept can easily be extended to intronic miRNAs that would target the transcription of their own hosts and their regulating elements, such as transcription factors or enhancers (Figure 3), as speculated by Li et al. [93]. However, to the best of our knowledge, there have been no large-scale experiments hypothesizing either complementary action or negative feedback regulation for intronic miRNAs.

The hypothesis tested in this study is that the key aspect of understanding the role of intragenic miRNAs lies in the functional relationship to their host genes. The integration of diverse data sources (Figure 4) is necessary for multi-dimensional investigation of the unique role of intragenic miRNAs. By analyzing the cellular role of their host genes, the introns they reside in and their targets, it is possible to build a comprehensive picture of the impact of intragenic miRNAs on cellular processes, potentially constituting a new approach to understanding carcinogenesis.

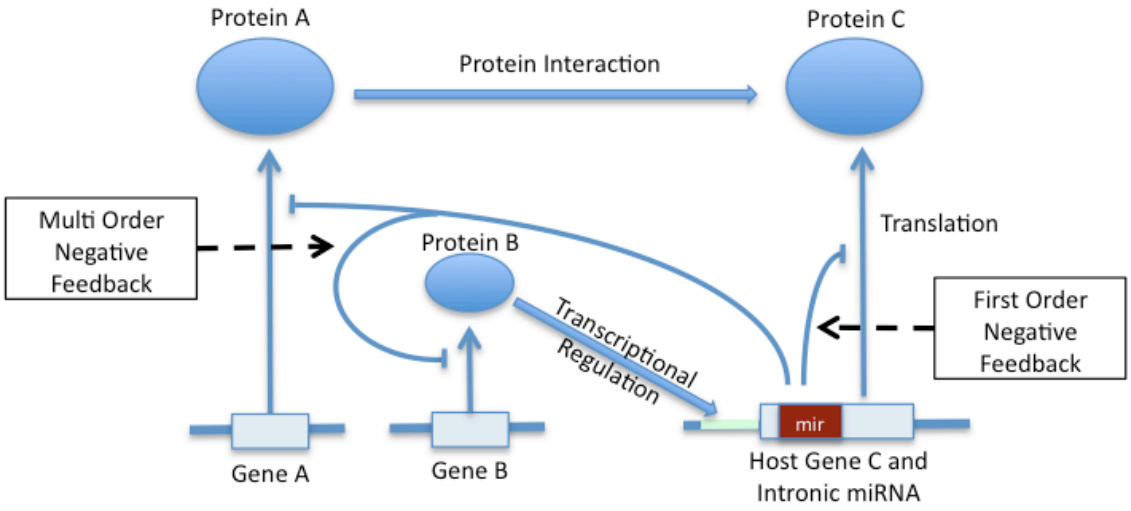

Figure 3 - Negative Feedback

Intragenic miRNAs could act as negatively regulating elements on multiple levels, by either targeting their host (referred to as “first order negative feedback”), regulating elements (such as transcription factors or enhancers) or proteins that interact with the protein expressed from the host gene (“multi order negative feedback”).

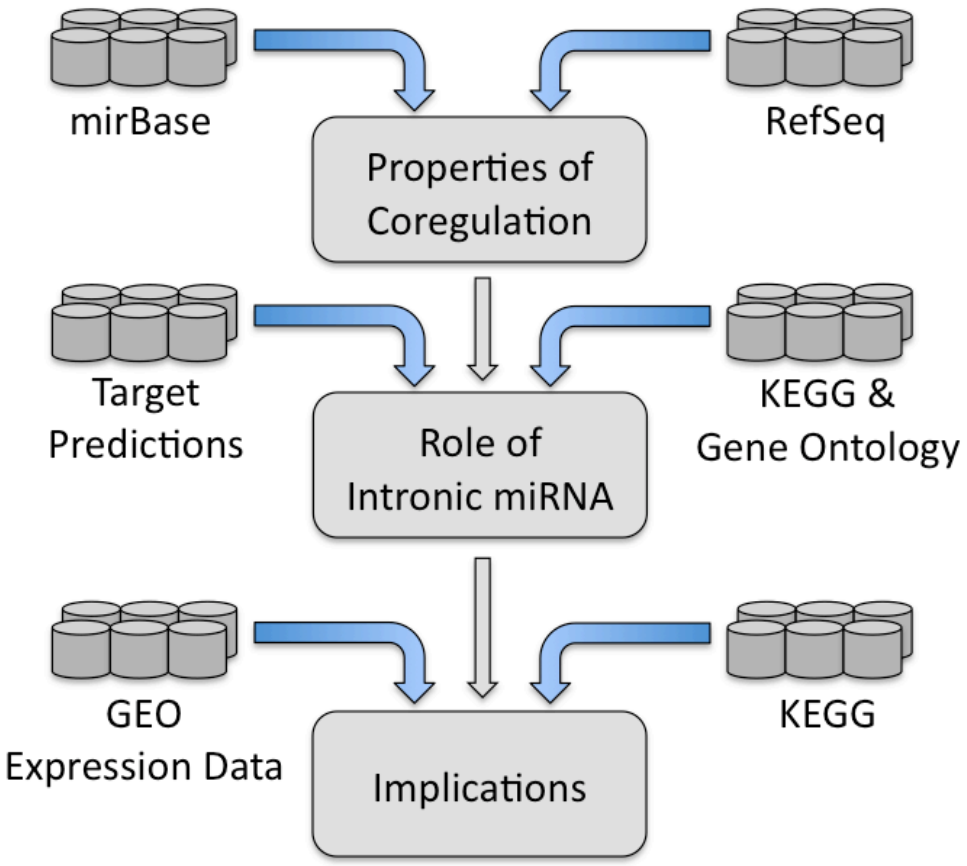

Figure 4 - Integration of Multiple Databases

The integration of multiple databases allows assessment of properties and functional aspect of miRNA – host co-regulation.

3 Materials and Methods

3.1 Public Datasets from GEO

3.1.1 Prostate Cancer (GSE7055)

Prueitt and co-workers performed a mRNA (Affymetrix HG-U133A 2.0) and miRNA (Ohio State University Comprehensive Cancer Center, Version 2.0) microarray expression analysis of samples from 57 patients with adenocarcinoma of the prostate [101]. Fifty of these showed perineural invasion (PNI), whereas 7 did not. None of the patients had undergone therapy prior to resection of the tumorous tissue. In addition to the microarray analysis, quantitative Real-Time PCR analysis was used to confirm measurements. Protein expression levels were assessed by immunohistochemistry. The authors found 19 miRNAs and 34 protein-coding genes to be differentially expressed between tumors with perineural invasion and those without (False Discovery Rate < 10%). All non-PNI tumors clustered together, with a subset of the PNI tumors from hierarchical clustering in gene ontology biological process (GOBP) analysis revealing statistical overrepresentation of differentially expressed genes in processes such as metabolism and transport of fatty, organic, amino acids and polyamines and processes related to negative regulation of programmed cell death.

3.1.2 Prostate Cancer mRNA (GSE6956)

Wallace et al. [102] hypothesized that differences in prevalence and lethality of prostate cancer in African-American and Caucasian-American men were due to differences in the tumor microenvironment. Therefore, they assessed the mRNA gene expression levels (Affymetrix HG-U133 2.0) of samples from 69 fresh frozen prostate adenocarcinomas (33 African-American men, 36 from Caucasian men) collected during 2002 – 2004. The tumors were all untreated and the presence of tumor tissue was confirmed by a pathologist. Eighteen non-tumor surrounding tissue samples were collected as negative controls. The authors were able to detect 162 transcripts that were differentially expressed between the two ethnic groups. In a disease association analysis, they related the identified transcripts to processes of autoimmunity and inflammation. Additionally, the

authors were able to build a 2-gene classifier to successfully distinguish between samples from each group.

3.1.3 Lung Adenocarcinoma mRNA (GSE7670)

Su et al. [103] suggested usage of *DDX5* as a novel internal control for quantitative real time polymerase chain reaction (Q-RT-PCR), to facilitate internal control evaluation and selection to corroborate microarray data. They used a dataset consisting of 66 lung samples. These included 27 cancer samples and surrounding normal tissue from patients at Taipei Veterans General Hospital, two tissue mixtures from Taichung Veterans General Hospital, two commercial human normal lung tissue samples, as well as epithelial and lung cancer cell lines. For the analysis presented here, only adenocarcinoma samples and cell lines as well as normal tissue samples were considered, resulting in 31 cancer samples and 29 normal controls.

3.2 Platforms

As mentioned earlier, microarrays are the basis for most contemporary investigations in miRNA expression levels in the cell. However, in contrast to mRNA platforms, there are currently few commercially available miRNA platforms, so many laboratories employ their own single-color cDNA spotted arrays. Due to the different nature of platforms, raw data was used for analysis.

3.2.1 Normalization of miRNA Datasets

Very little is known about how to best preprocess miRNA microarray data. Even though evidence suggests that quantile normalization might be the best method [104], there are no systematic studies of which background correction or summarization method works best for miRNA microarrays. In this study, the density plots of multiple current background correction methods were compared to no background correction in the "Prostate Cancer" dataset, and the results are summarized in Figure 5.

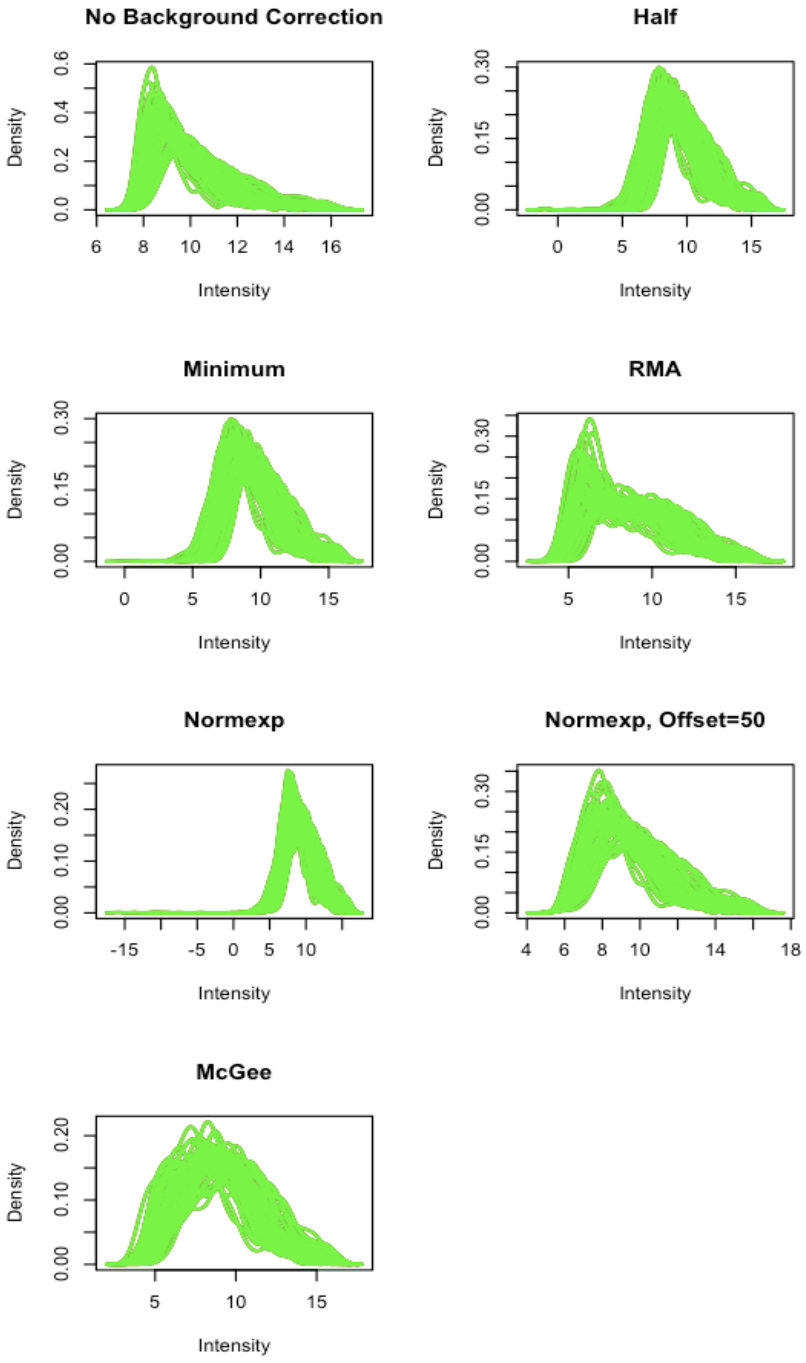

Figure 5 - Comparing Background Correction Methods

Several background correction techniques are compared to no background correction at baseline. Background subtraction, although popular, was not considered because of missing values.
The background correction methods “Half” and “Minimum” show distributions close to normal, which is helpful for detection of differentially expressed genes. Robust multi-array (RMA) expression measure [70, 72, 105], Normexp [69] and McGee [106] appear less optimal choices (all of the above methods are reviewed in [69]).

After background correction using the minimum of foreground and background intensity values, quantile normalization was used [70, 104]. The resulting plots before and after normalization are shown in Figure 6 and Figure 7, respectively.

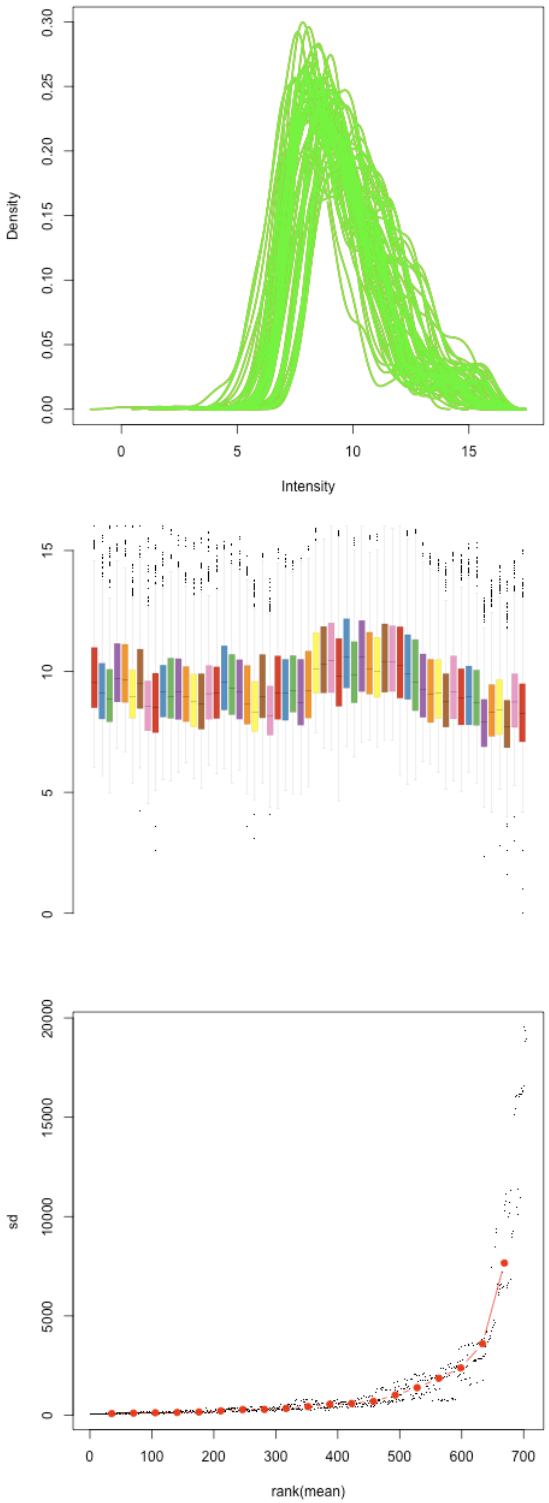

Figure 6 - Expression Values Before Normalization

The top graphic shows a plot of intensity values before normalization. The box plot (middle) visualizes mean and standard deviation of the individual microarrays. The bottom graphic shows dependence of variance on mean intensity.

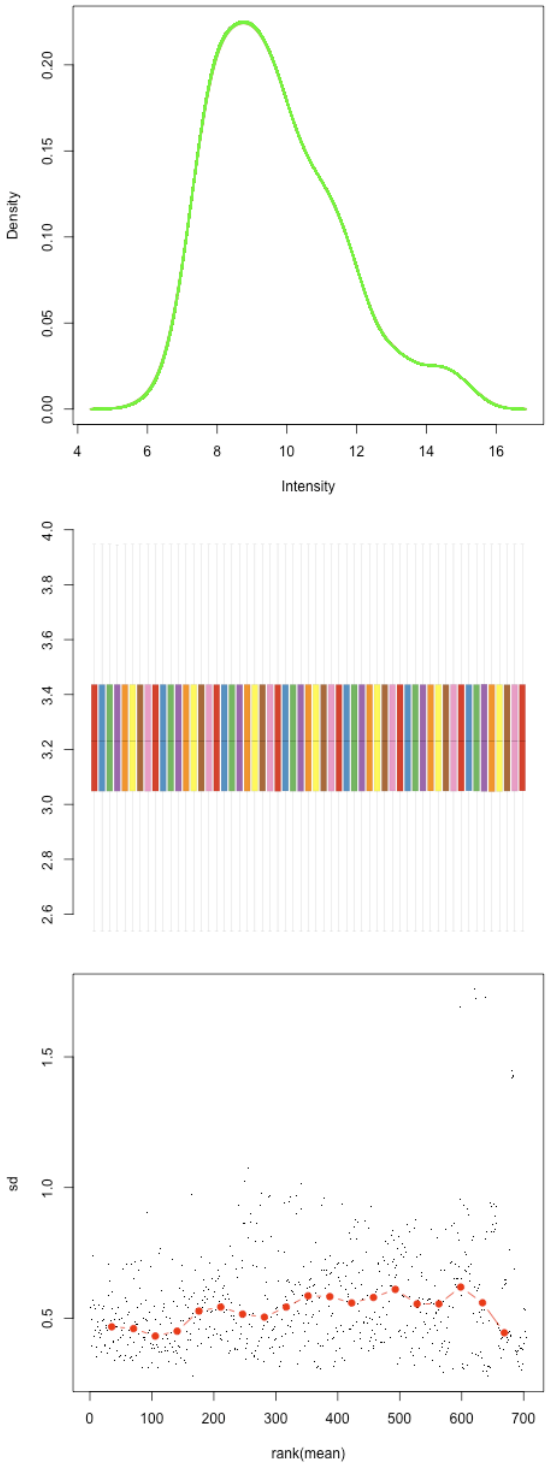

Figure 7 - Expression Values After Normalization

After normalization, intensities appear to be normally distributed (top), all the arrays seem to have the same mean and standard deviation (middle) and the variance seems to be independent of the mean intensity (bottom).

3.3 Databases and Classification of miRNAs

3.3.1 Design of the database

In order to use a common nomenclature, gene info files from the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/), including the fields "gene symbol", "gene name", "Ensembl identifier" and "synonyms", were imported to a local database. Also imported was the NCBI RNA reference sequences collection (RefSeq) release 31 [107] from University of California, Santa Cruz (UCSC), matching those entries to a gene symbol or its synonym in the database. In full, this totaled 27,235 entries representing 18,684 distinct genes. Exon start and end coordinates were also imported (291,478 entries), and (+1) was added to every start coordinate, as described on the UCSC website (http://genome.ucsc.edu).

miRBase release 11.0 (April 2008) [108-112] contains current information on known miRNAs in different organisms, including human, mouse, chicken, dog, worm, and zebrafish, providing the official miRNA symbol as well as genomic coordinates. A summary of the number of known microRNAs of organisms imported for use in this study is shown in Table 2.

Table 2 – Known Number of Distinct miRNAs for Different Species (miRBase release 11.0, April 2008)

Organism	Number of miRNAs	Number of Known Genes (RefSeq)	Ratio Protein Coding Genes : miRNAs
Homo sapiens	692	18693	27:1
Mus musculus	482	19228	40:1
Canis familiaris	204	912	4:1
Gallus gallus	469	4158	9:1
Danio rerio	318	13204	42:1
Drosophila melanogaster	152	14072	93:1
Caenorhabditis elegans	154	19612	127:1

The genomic position of the miRNAs were mapped to known protein coding genes registered in RefSeq, to identify intragenic miRNA whose genomic position lay within the

transcription start and the transcription end position of an annotated gene ("host gene"). Subsequently, intragenic miRNAs were further subdivided into intronic and exonic. An intragenic miRNA was labeled exonic, if its genomic coordinates overlapped with genomic coordinates of any exon in the database, and was labeled intronic otherwise. In addition, intragenic miRNAs can be classified depending on whether they are on the same or the opposite strand of their host gene. In cases where a miRNA overlapped with two different genes on two strands, the gene on the same strand was considered the host gene. This choice, however, affected only few entries. If the miRNA position overlapped with two genes on the same strand, the larger gene was selected. The distance to the next upstream exon and the intron length, as defined by the region between the immediate upstream and downstream exon, were also calculated. The distributions of intronic, exonic and intergenic genes for different organisms are shown in Table 3. The distribution of strand direction for intronic miRNAs and their host genes is shown in Table 4 (Note: The row sums of Table 3 are greater than the number of distinct known miRNAs in Table 2 because different copies of the same miRNAs may be double counted as intergenic and intragenic, as is the case for *hsa-mir-1184*).

Table 3 – Distribution of Classes of miRNAs

Organism	Intronic	Exonic	Intergenic
Homo sapiens	296 (42.6 %)	37 (5.3 %)	362 (52.1 %)
Mus musculus	171 (35.4 %)	30 (6.2 %)	282 (58.4 %)
Canis familiaris	3 (1.5 %)	0 (0 %)	201 (98.5 %)
Gallus gallus	50 (10.7 %)	1 (0.2 %)	418 (89.1 %)
Danio rerio	48 (15.0 %)	1 (0.3 %)	271 (84.7 %)
Drosophila melanogaster	65 (42.8 %)	2 (1.3 %)	85 (55.9 %)
Caenorhabditis elegans	51 (33.1 %)	1 (0.6 %)	102 (66.2 %)

Table 4 – Distribution of Direction of Intronic miRNAs with Respect to their Host Gene

Organism	Number of Intragenic miRNAs on the Same Strand as Host Gene	Number of Intragenic miRNAs on the Opposite Strand of Host Gene
Homo sapiens	282 (84.7 %)	51 (15.3 %)
Mus musculus	163 (78.2 %)	38 (21.8 %)
Canis familiaris	2 (66.7 %)	1 (33.3 %)
Gallus gallus	46 (90.2 %)	5 (9.8 %)
Danio rerio	39 (79.6 %)	10 (20.4 %)
Drosophila melanogaster	53 (79.1 %)	14 (20.9 %)
Caenorhabditis elegans	33 (63.6 %)	19 (36.5 %)

3.3.2 Calculation of Intron Position, Intron Size and Distance to Upstream Exon

RefSeq may contain multiple observations for a given gene, usually revealing distinct patterns of alternative splicing. Therefore, it is important to decide, based on the nature of the question and underlying biological assumptions, when a region will be called an exon or an intron, given that there is evidence of both. In all the experiments, a region was considered an exon if and only if there was at least one RefSeq identifier for which this region was labeled exonic. All overlapping exons were merged into one exonic region.

3.3.3 Calculating an Expected Distribution of Introns

The expected proportion of miRNAs in a given intron was calculated as follows:

Assuming an equal chance for a miRNA to be in any intron of a host gene, for intron j in gene i and n total introns: $p_{i,j} = \frac{1}{n}$, for $j \leq n$ and 0 otherwise. The proportion of miRNAs

in a given intron j can be calculated by the normalized weighted sum of probabilities $p(j) = \frac{\sum_{i=1}^{m} p_{i,j}}{m}$, where m is the total number of genes considered. This allows the estimation of the expected number of miRNAs, by multiplication with the total number of hosts. An example is visualized in Figure 8.

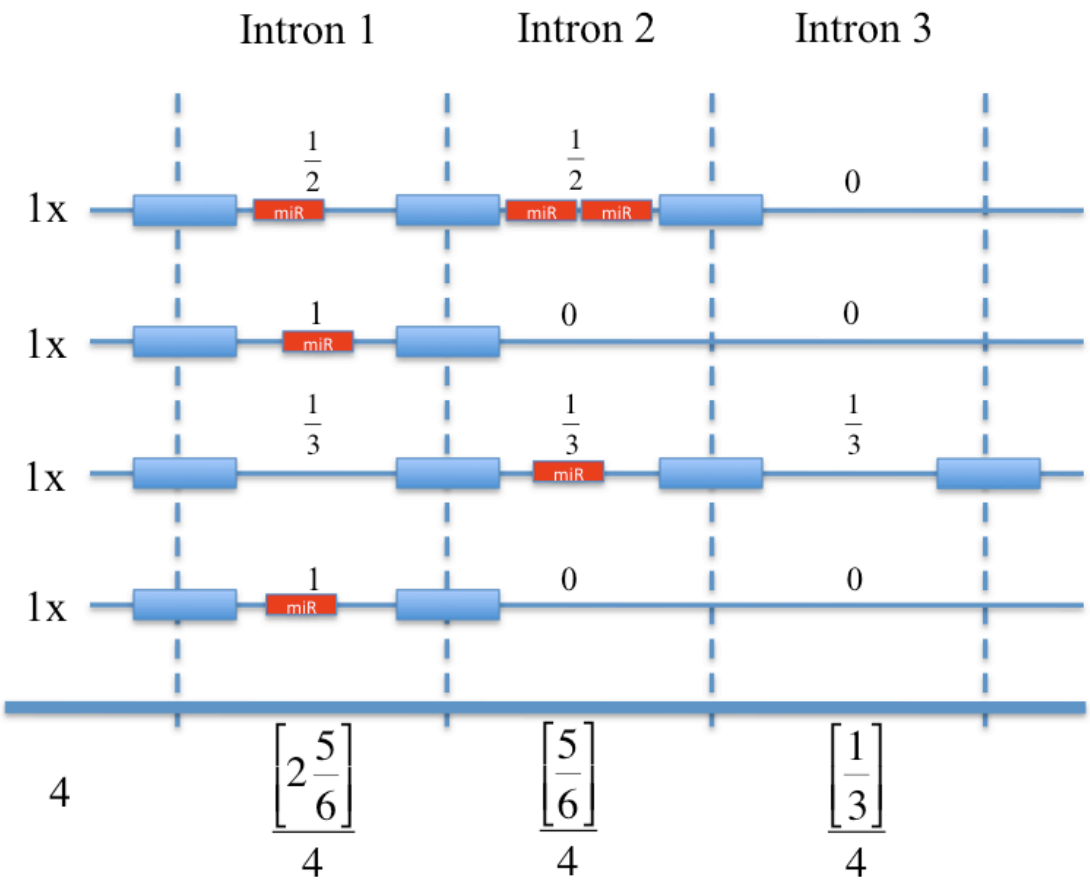

Figure 8 - Calculating the Expected Intron Distribution

The expected probability that a miRNA occurs in a certain intron can be calculated by normalizing the weighted sum of probabilities for individual introns.

3.4 Target Prediction Methods

3.4.1 Import of Target Predictions

Precalculated target predictions for TargetScan release 4.2 [77] (April 2008), PITA (top 15%) [84] catalog version 6 (August 2008), MirTarget2 (mirDB) version 2.0 [35] (December 2007), miRanda [80] (September 2008), RNA22 [81] (November 2006) and

PicTar 5-way [86] were downloaded. Also included was TarBase version 5.0c [89] (June 2008) as a reference database for miRNA target interactions with published evidence. Only targets that had an assigned value of either "True" (which typically means experimental validation via luciferase reporter assay) or "Microarray" for the variable "Support Type" were selected.

Some miRNA symbols did not exactly match entries in the database for various reasons, including use of non-official names or older miRBase releases. Whenever a miRNA symbol could not be found, matching was attempted to an extension such as "-1" or "a" (for example, *hsa-mir-511* in mirTarget2 was matched to *hsa-mir-511-1* and *hsa-mir-511-2*). If the miRNA symbol ended with a letter, it was removed to check for other matches (from the PicTar prediction list *hsa-mir-128a* matched to *hsa-mir-128-1*, *hsa-mir-128-2*, and *hsa-mir-128-3* for example). Predictions for a miRNA symbol were ignored if no matches could be found.

3.5 Pathway Analyses

3.5.1 Gene Ontology

The Gene Ontology (GO) [113] classifications of all 246 host genes of intragenic miRNA genes that were located on the same strand as their host gene were surveyed using Cytoscape 2.6.0 [114] and BiNGO 2.3 [115]. We focused our attention on those categories that were disproportionately overrepresented. The setting "Hypergeometric test" was chosen to calculate the probability of observing an equal or greater number of genes in a given functional category that is shared among n genes of the reference set (consisting of all known genes) than in the test set x. The False Discovery Rate (FDR), which is the standard setting in BiNGO 2.3 [115], was controlled.

3.5.2 Statistical Software

The statistical programming software R 2.7.1 [116] was used in combination with bioconductor [117] packages AnnBuilder 1.18.0 [118], KEGG.db version 2.2.0 and GOStats version 1.7.4 [119] to acquire a list of pathways that were associated with one or more of the 246 host proteins.

3.5.3 Target Coverage

The union of predicted targets included more than 90% of all known human genes. Since target prediction methods are very different, they are difficult to compare. In this work, only targets that were predicted by at least two different methods were considered in the calculation of target coverage. This reduced the total number of predictions by almost 70%, as can be seen in Figure 9.

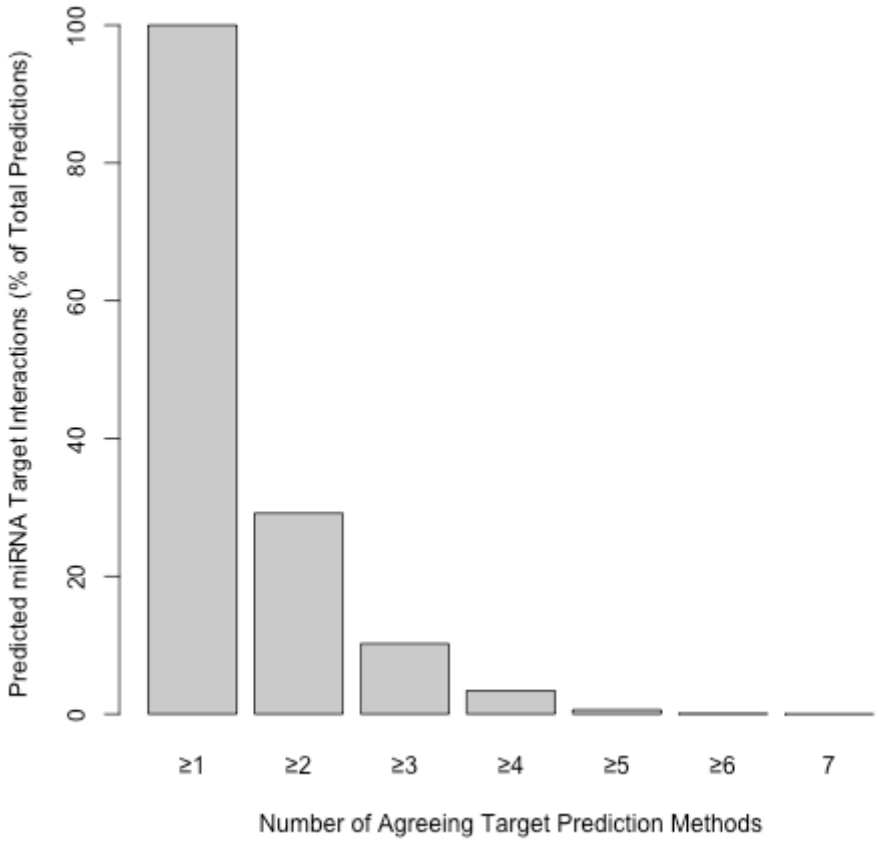

Figure 9 – Target Prediction Agreement of Different Methods

The requirement that two different methods had to agree on a target prediction for a given (intronic) miRNA reduced the total number of predictions by almost 70%.

We defined the set S_p as the set of genes linked to a pathway and S_t as the set of predicted targets of the miRNAs associated with the pathway through their host genes. The target coverage (C) for a pathway was defined as

$$C = \frac{|S_p \cap S_t|}{|S_p|} .$$

Statistical significance of target enrichment within a pathway was tested by randomly sampling $|S_p|$ genes from a universe of all known genes, replacing the genes within the pathway with the set of genes in the random sample (S_i), and subsequently calculate a new "random" target coverage C_i'. This procedure was repeated 1000 times, allowing to estimate the probability as the number of times a target coverage C_i' greater or equal to C was observed. We defined the indicator function I(C_i',C) as

$$I(C_i', C)\begin{cases} 1 & if\ C_i' \geq C \\ 0 & otherwise \end{cases} .$$

Hence, the probability of observing a greater or equal target coverage for a given pathway can be estimated as

$$p(C' \geq C) = \frac{\sum_{i=1}^{1000} I\left(\frac{|S_i \cap S_t|}{|S_i|}, C\right)}{1000} \text{ , where } |S_i| = |S_p| .$$

Analogously, the enrichment statistics for miRNAs targeting their own hosts were calculated, where S_p was defined as the set host genes, S_t as the set of targets of the intragenic miRNAs of these host genes and S_i as the set of $|S_p|$ randomly sampled genes from the universe of all predicted targets for these miRNAs.

4 Results

4.1 Intronic miRNAs

4.1.1 Intronic miRNAs Have a Positional Bias Towards 5′ Introns

The orientation of the gene for an intronic miRNA depends significantly on the direction of its host strand ($p = 1.3 \times 10^{-36}$ in X^2 test) as shown in Table 4. This feature is thought to be beneficial to the cell [7]. However, the distribution of miRNAs across their hosts' introns might as well be of functional significance. Therefore, the distribution of introns containing miRNA genes was compared to an expected distribution calculated by assuming an equal probability of occurrence for each intron of a host gene. Interestingly, it seems that intronic miRNA genes have a positional bias towards the early 5' introns (Figure 10), when compared to the expected distribution (p-value = 0.02 in X^2 test).

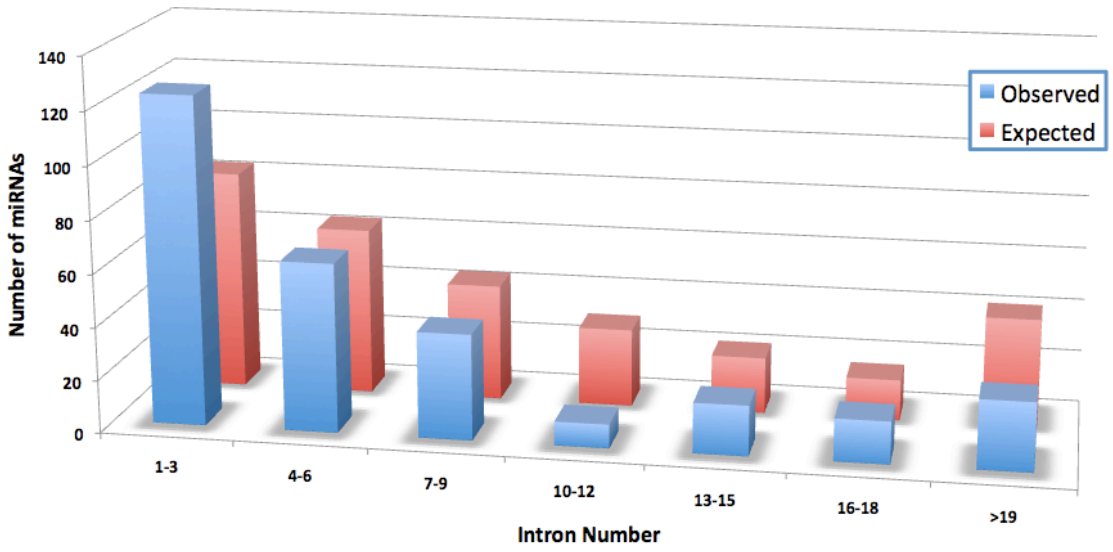

Figure 10 - Distribution of miRNAs Across Introns of their Host Genes

Intronic miRNAs seem to have a positional bias towards introns closer to the 5' end ($p = 0.02$ in X^2 test).

It is well known that transcriptional activity is higher towards the 5' region of a gene [120] and also that regulatory motifs tend to reside in these regions [92]. This finding supports the idea of a functional linkage between host gene and miRNA.

4.1.2 Reduced Host – miRNA Correlation in Cancer Samples

miRNA and mRNA expression of 57 prostate cancer samples that were previously published [101] were compared. For 42 of the potential 331 [miRNA – host] pairs, correlation coefficients and their significance level could be calculated. From the 42 pairs, 35 miRNAs were on the same strand as their host, and 7 were located on the opposite strand. The average Pearson correlation coefficient was +0.12, which was significantly higher than would be expected by chance ($p<0.001$ in 1,000 random permutations). However, in contrast to Baskerville and Bartel [7], who found that 67% of miRNAs displayed a higher absolute correlation with their hosts than with up- or downstream genes, in this analysis only 20% ($p<0.05$) were found to be significantly correlated, supporting findings of [95].

Independent regulation of intronic miRNAs has been hypothesized for large introns, implying the existence of a potential regulatory region within the intron. This idea could also been seen in the context of the finding that intronic miRNAs have a bias towards the 5'-introns, which are believed to contain regulatory regions [92]. Figure 11 displays the relationship between the miRNA expression and host mRNA expression as a function of the host intron size. Figure 12 displays the relationship between the miRNA expression and host mRNA expression as a function of the distance to the next exon upstream. Figure 13 displays the relationship between the miRNA expression and host mRNA expression as a function of the intron number. It appears that large intron size decreases the correlation of miRNA and host mRNA expression, and that distance to the upstream exon has the same effect.

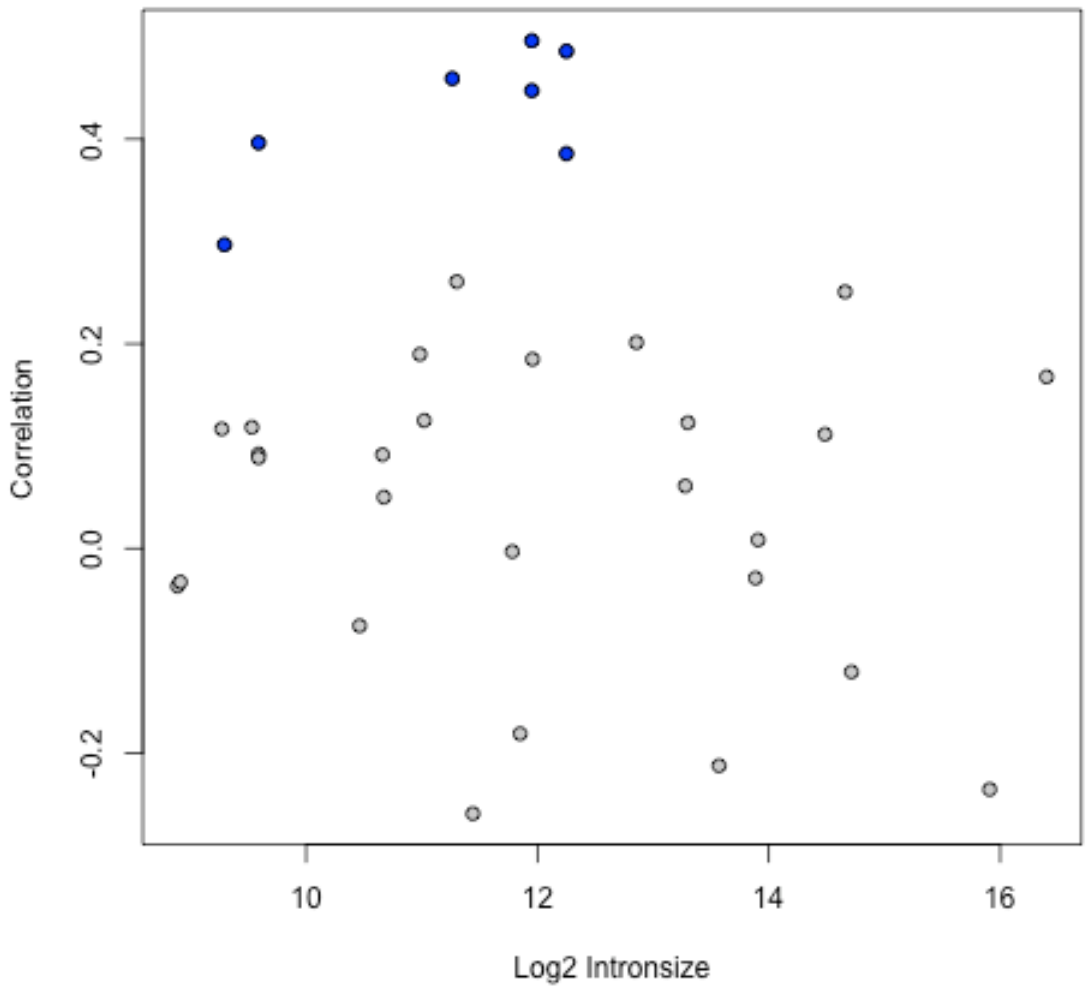

Figure 11 - Correlation of Expression of miRNA and Host (Intron Size)

When the correlation between intronic miRNA expression and the expression of its host is visualized according to the size of the corresponding intron, significant correlation ($p < 0.05$) is only observed up to a total intron size of 4-8kb.

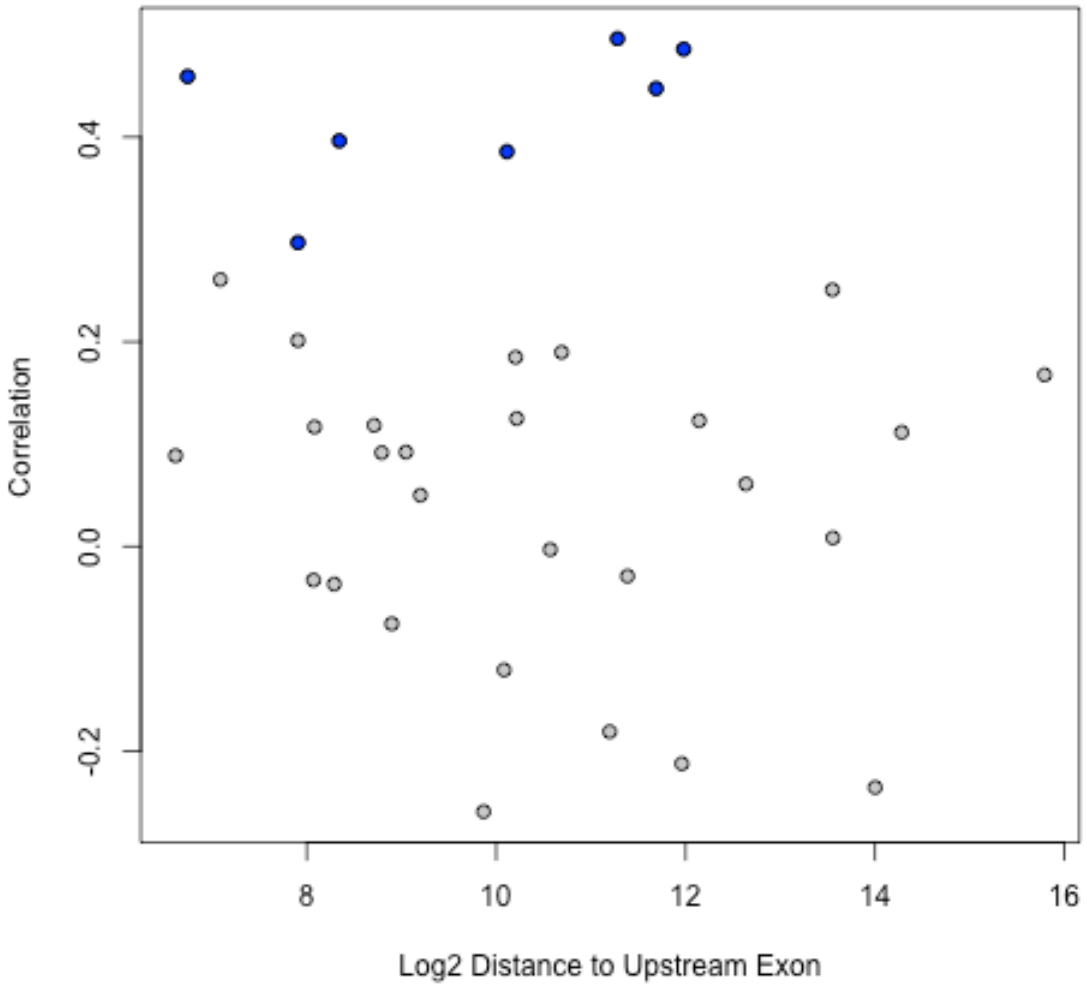

Figure 12 - Correlation of Expression of miRNA and Host (Distance to Exon)

Similar results as in Figure 11 are observed when looking at correlation of host expression and intronic miRNA expression according to the distance to the closest upstream exon. Again, no significant correlation ($p < 0.05$) is observed for a distance of greater than 4 – 8 kb.

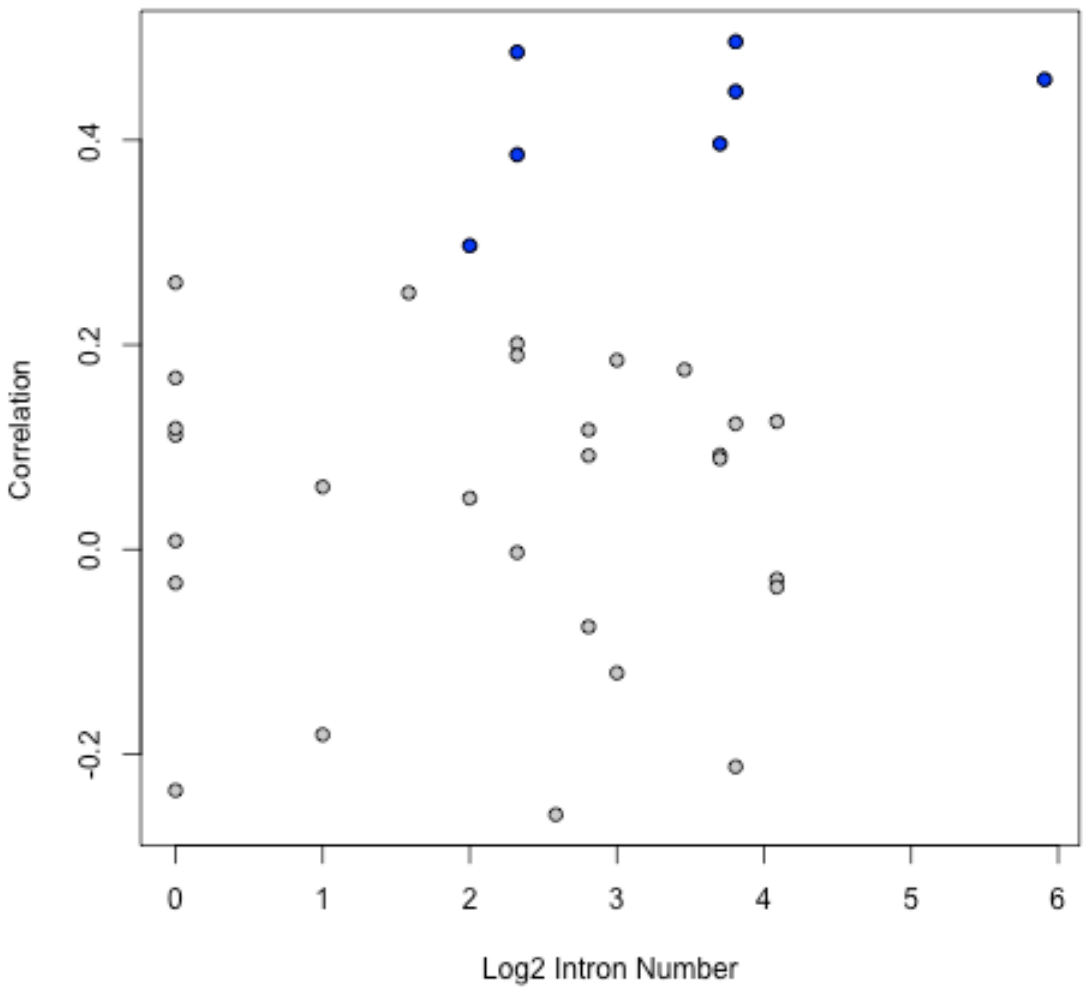

Figure 13 - Correlation of Expression of miRNA and Host (Intron Number)

Correlation seems to be independent of intron number (intron #1 is defined as the one closest to the TSS of the gene).

4.1.3 Intragenic miRNAs Target Their Hosts

In the recent past, different roles for intronic miRNAs have been claimed. Whereas Li et al. attributed the major significance of the relationship between these molecules and their host genes to a negative feedback regulatory mechanism [93], Barik identified an intronic miRNA that would target genes functionally antagonistic to its host [99]. In a recent

report on computational miRNA prediction in amphioxus, Luo and Zhang reported that intronic miRNAs do not have complementary target sites in their host genes, but in neighboring genes [121], which may suggest a multi-order negative feedback.

Sixty-one miRNAs that potentially target their host genes (predicted by at least one method) were identified in our computational analyses, corresponding to 53 different host genes. By exchanging the set of host gene names for a set of randomly sampled gene names, this number is shown to be significantly higher than expected by chance ($p < 0.01$). The background distribution is shown in Figure 14. This result strongly supports the idea that intronic miRNAs can potentially act as first-order negative feedback regulators.

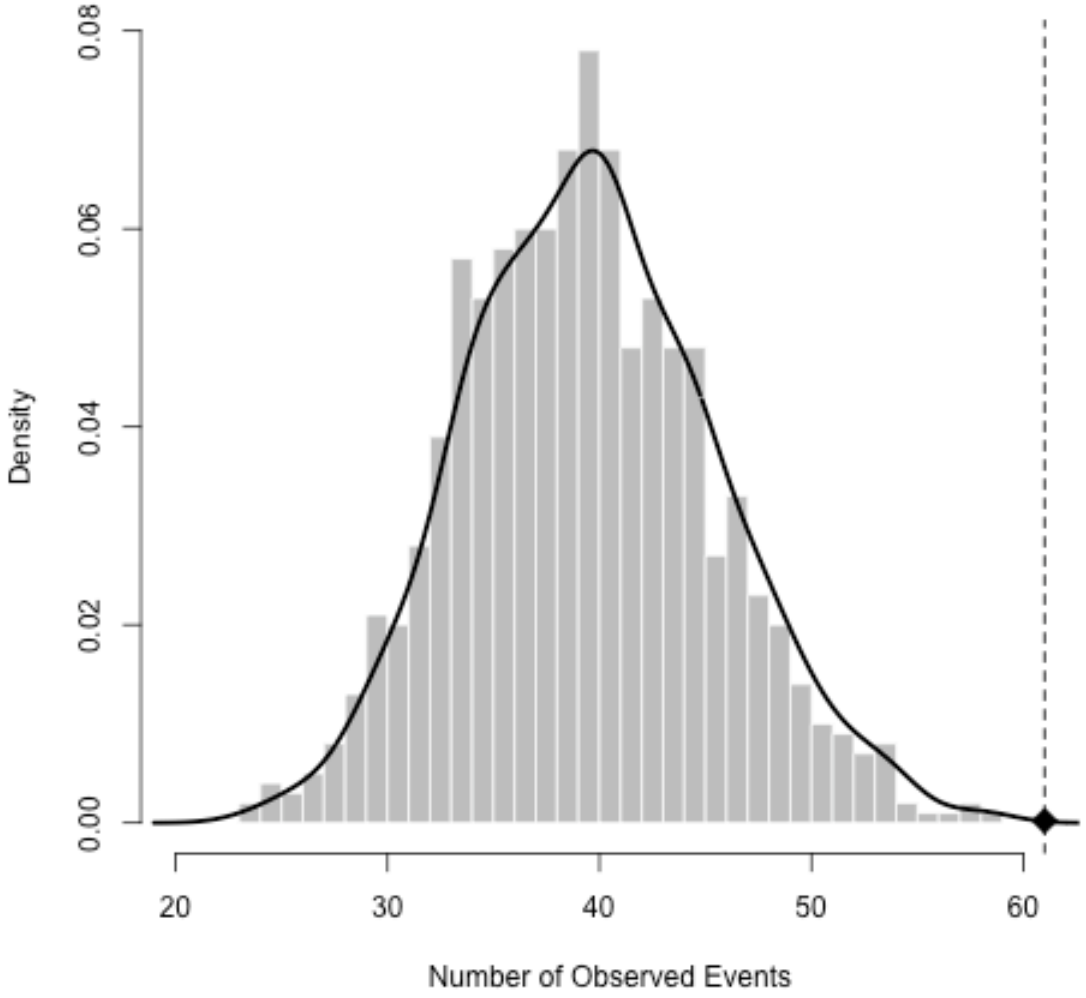

Figure 14 - Intronic miRNAs Targeting Their Hosts

The background distribution derived from 1000 random samplings follows a Gaussian normal distribution (quantile-quantile plot not shown) with a mean of 39.7 and a standard deviation of 5.9. The probability of observing 61 miRNAs that target their own hosts can therefore be estimated to be $p = 6x10^{-10}$.

4.2 Functional Analysis of Host Proteins

4.2.1 Gene Ontology Analysis

Table 4 shows that most intragenic miRNA genes are read in the same direction as their host genes, suggesting a common involvement in biological processes. This supports the

idea of indirectly assessing functional aspects of intragenic miRNAs by looking at current knowledge about the host genes and their gene products. A GO analysis of the host genes was performed, looking for overrepresentation of host genes in the ontologies "biological processes", "molecular function", and "cellular component". A biological process is defined as being linked to a biological objective, such as "signal transduction" or "translation". It is comprised out of the interplay of multiple "molecular functions", which in turn are defined as specific roles of a protein, such as "enzyme", "ligand" or "adenylatcyclase". "Cellular component" describes the place of action of the gene product in the cell [113]. As has been reported previously [91], hosts of intragenic miRNAs are involved in a broad spectrum of cellular functions, the major ones including metabolism, biosynthesis and gene regulation (Figure 15, Figure 16). This is in accordance with the general notion that miRNAs are important regulators of cell development and interaction. Location-wise, main categories belong to synaptic processes, cell adherence, communication and muscle development (Figure 17).

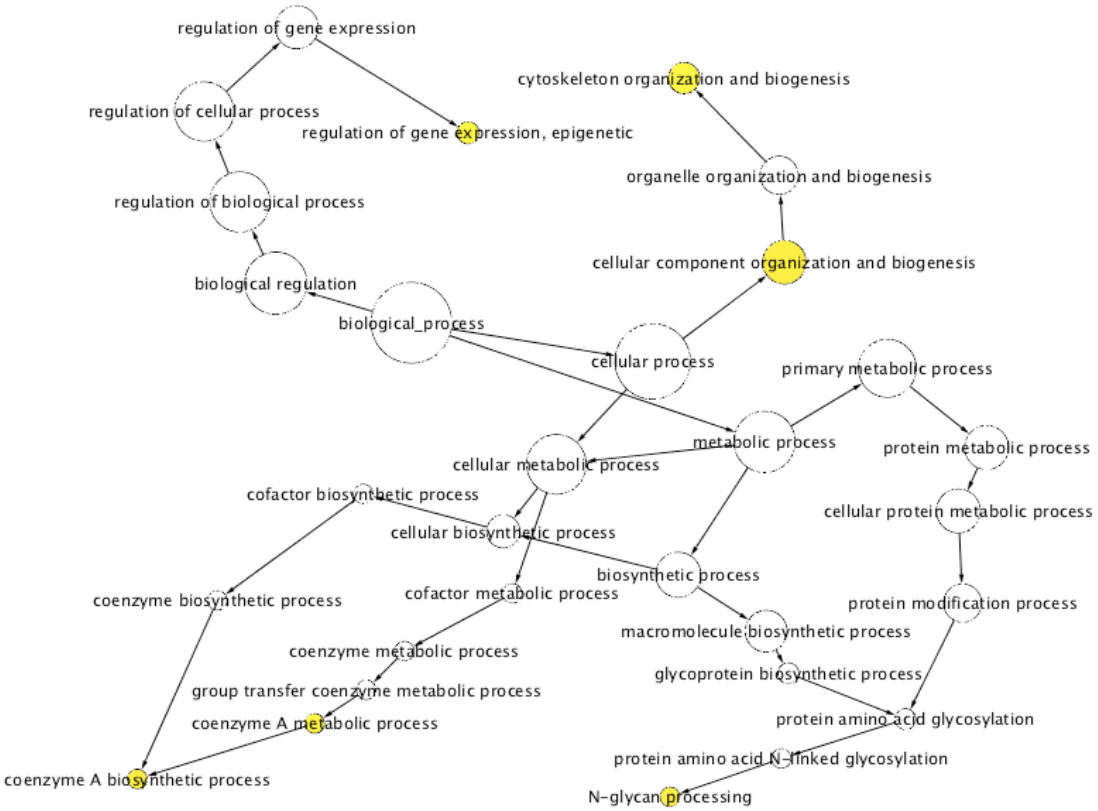

Figure 15 - Hosts in GO Biological Processes

Figure 15 shows overrepresentation of host genes in different categories of the “biological processes” ontology. A yellow node indicates statistical overrepresentation of host genes in the respective category.

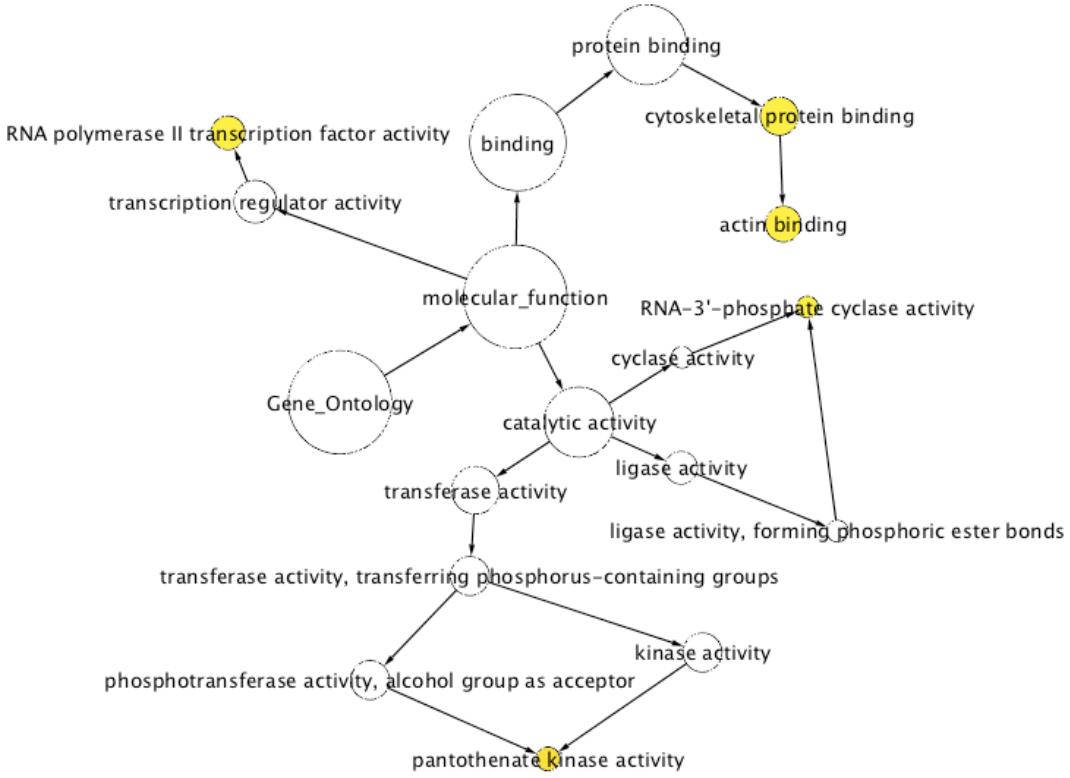

Figure 16 - Hosts in GO Molecular Function

In the ontology “molecular function”, five categories show significant overrepresentation, including the broad areas of transcriptional regulation, protein binding, and catalytic activity.

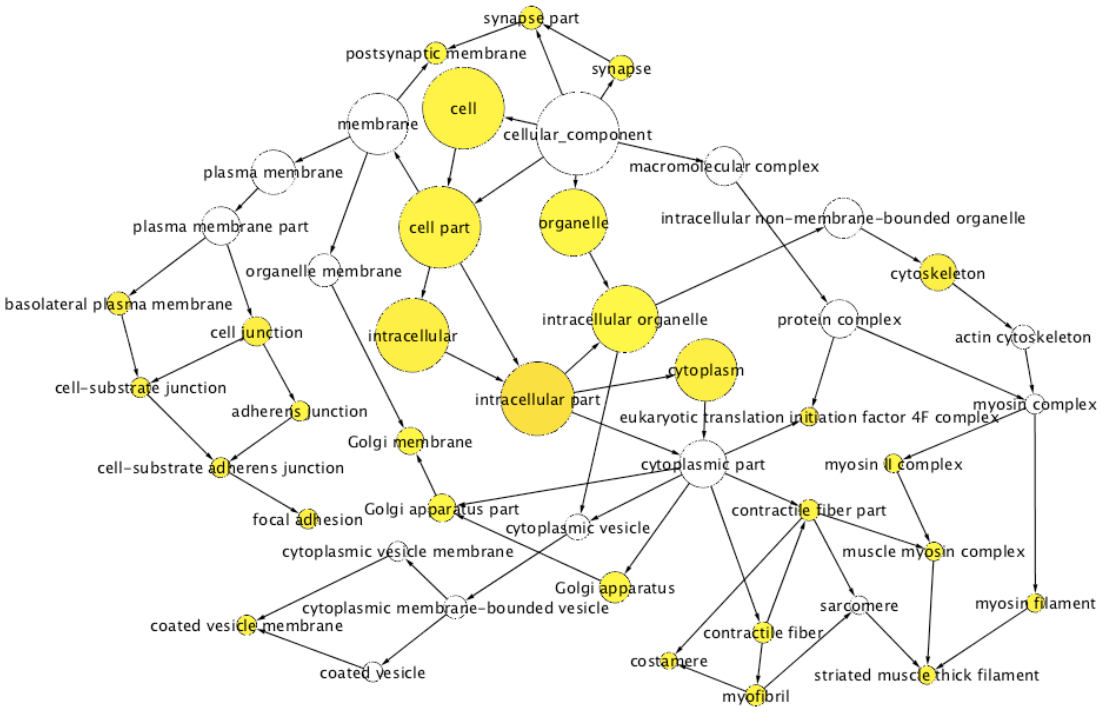

Figure 17 - Hosts in GO Cellular Component

An analysis of associated locations of hosts of miRNA genes indicates presence in many distinct parts of the cell. Prevailing categories include synaptic processes, cell adherence, and communication and muscle development.

4.2.2 KEGG Analysis Suggests Role of Hosts in Signaling Pathways

The "Kyoto Encyclopedia of Genes and Genomes" (KEGG) [122-124] is a collection of multiple databases. The KEGG Pathways database contains information on biochemical pathways and protein interactions, hence representing molecular interaction networks, including metabolism, genetic information processing, environmental information processing, cellular processes, human diseases and drug development. Due to the nature of the database, statistical analyses can be performed equivalently to those in GO, and the results are summarized in Table 5. Hosts of intragenic miRNAs are significantly overrepresented in twelve pathways ($p < 0.05$). The majority of significant pathways are involved in signaling processes (MAPK signaling, axon guidance, ErbB signaling, VEGF signaling, calcium signaling), followed by biosynthetic processes (panthothenate and CoA biosynthesis, glycan structures biosynthesis, biosynthesis of fatty acids).

Table 5 - Overrepresentation of Hosts in KEGG Pathways

Pathway	Expected Number of Host Genes in Pathway	Observed Number of Host Genes in Pathway	Total Number of Genes in Pathway	p-Value
MAPK Signaling	4	11	264	0.001
Pantothenate & CoA Biosynthesis	0	3	16	0.001
Axon Guidance	2	7	128	0.002
ErbB Signaling	1	5	87	0.008
Tight Junction	2	6	135	0.013
DRPLA	0	2	15	0.019
VEGF Signaling	1	4	73	0.021
Type 1 Diabetes Mellitus	1	3	43	0.024
Neuroactive Ligand-Receptor Interaction	4	8	255	0.030
Glycan Structures – Biosynthesis	2	5	122	0.031
Calcium Signaling	3	6	176	0.041
Biosynthesis of Unsaturated Fatty Acids	0	2	23	0.043

However, statistical over-representation of host genes of intronic miRNAs is not the only interesting feature. The full list of pathways itself presents an interesting inside into the spectrum of functional association of these genes. Interestingly, intragenic miRNAs are present in 16 out of the 21 KEGG signaling pathways, some of which have been shown to play a prominent role in carcinogenesis, like MAPK signaling [125], Erbb signaling [126], Calcium signaling [127], and mTor signaling [128].

4.2.3 Intronic miRNAs Target Multiple Genes in Their Hosts' Pathways

In order to test the hypothesis that intronic miRNAs might act as regulators in the global context of a negative feedback loop circuitry, the KEGG pathway analysis was extended to identify targets within the biomolecular pathway. To understand the trade-off of sensitivity and specificity in existing target prediction algorithms, the number of agreed targets was plotted against the number of algorithms in which that prediction was made (see Figure 18). In order to check whether the observed target coverage was expected by chance, the original genes contained in the pathway were replaced by a set of randomly sampled genes and the expected target coverage was calculated. The distributions of expected target coverages are visualized in Figure 19 and Figure 20.

When a prediction agreement of ≥ 2 methods was required, 25 pathways out of 74 had a significant overrepresentation of targets at a threshold of 0.05 (Table 6).

Even though there is significant overlap between Table 5 (overrepresentation of hosts) and Table 6 (overrepresentation of targets), it is interesting to observe that in cancer pathways are ranked high especially among pathways in Table 6.

Table 6 - Target Overrepresentation in KEGG Pathways

Pathway	Host Genes	Target Coverage	p-Value
Axon Guidance	PPP3CA, PTK2, SEMA4G, SEMA3F, SLIT2, SLIT3, ABLIM2	33.6%	< 0.001
ErbB Signaling	ERBB4, AKT2, PRKCA, PTK2, MAP2K4	32.2%	< 0.001
Long-term Potentiation	PPP3CA, PRKCA, RPS6KA2	18.6%	< 0.001
MAPK Signaling	ATF2, DDIT3, AKT2, FGF13, ARRB1, PPP3CA, PRKCA, CACNG8, RPS6KA2, MAP2K4, RPS6KA4	30.3%	0.001
Focal Adhesion	COL3A1, AKT2, PRKCA, PTK2, TLN2	25.8%	0.001
Non-Small Cell Lung Cancer	AKT2, PRKCA	25.9%	0.001
Glioma	AKT2, PRKCA	27.7%	0.001
Pancreatic Cancer	AKT2	19.2%	0.001
Regulation of Actin Cytoskeleton	CHRM2, FGF13, SSH1, PTK2	17.6	0.003
Melanogenesis	PRKCA	10.78%	0.003
Tight Junction	AKT2, MYH6, MYH7, PRKCA, ASH1L, MYH7B	25.9%	0.004
Bladder Cancer	DAPK3	19.0%	0.004
Prostate Cancer	AKT2	18.0%	0.004
T Cell Receptor Signaling	AKT2, PPP3CA	16.1%	0.005
Amyotrophic Lateral	PPP3CA	10.5%	0.007

Sclerosis			
Colorectal Cancer	AKT2	16.7%	0.007
GnRH Signaling	PRKCA, MAP2K4	10%	0.01
Calcium Signaling	CHRM2, ERBB4, HTR2C, ATP2B2, PPP3CA, PRKCA	25.6%	0.012
Ubiquitin Mediated Proteolysis	HUWE1, WWP2, BIRC6, ITCH	23.9%	0.022
Melanoma	AKT2, FGF13	19.7%	0.022
Insulin Signaling	AKT2, SREBF1	15.8%	0.034
Cell Cycle	MCM7	24.1%	0.037
Chronic Myeloid Leukemia	AKT2	14.5%	0.039
Glycan Structures Biosynthesis	MGAT4B, FUT8, CSGLCA-T, GALNT10, HS3ST3A1	18.9%	0.045
Small Cell Lung Cancer	AKT2, PTK2	16.1%	0.048
Apoptosis	AKT2, PPP3CA	14.3%	0.052
FC epsilon RI Signaling	AKT2, PRKCA, MAP2K4	16.9%	0.091

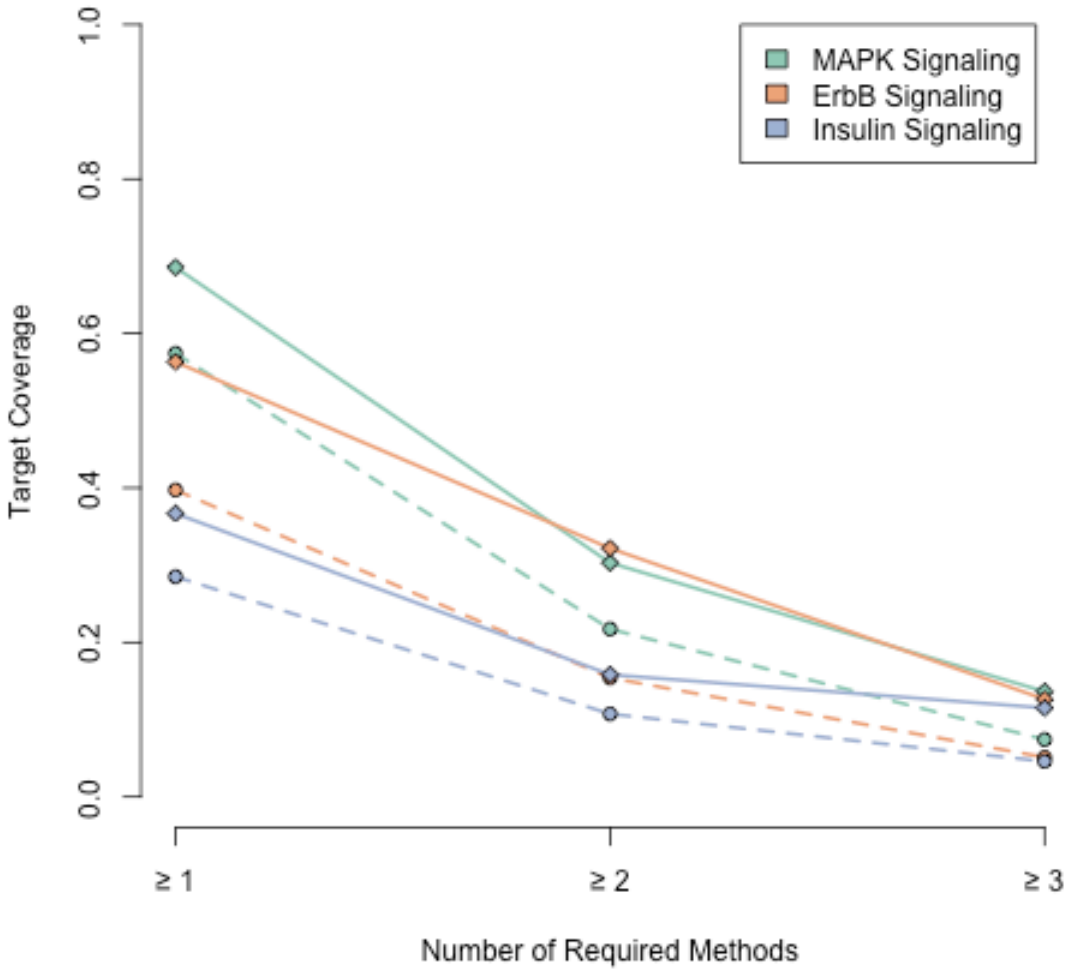

Figure 18 – Influence of Prediction Agreement on Target Coverage

Increasing the required agreement between the different prediction methods increases specificity and decreases sensitivity. Solid lines represent observed target coverage, dashed lines indicate the by chance expected target coverage. The difference between a solid and dashed line is an estimate of the relationship between the underlying signal and noise.

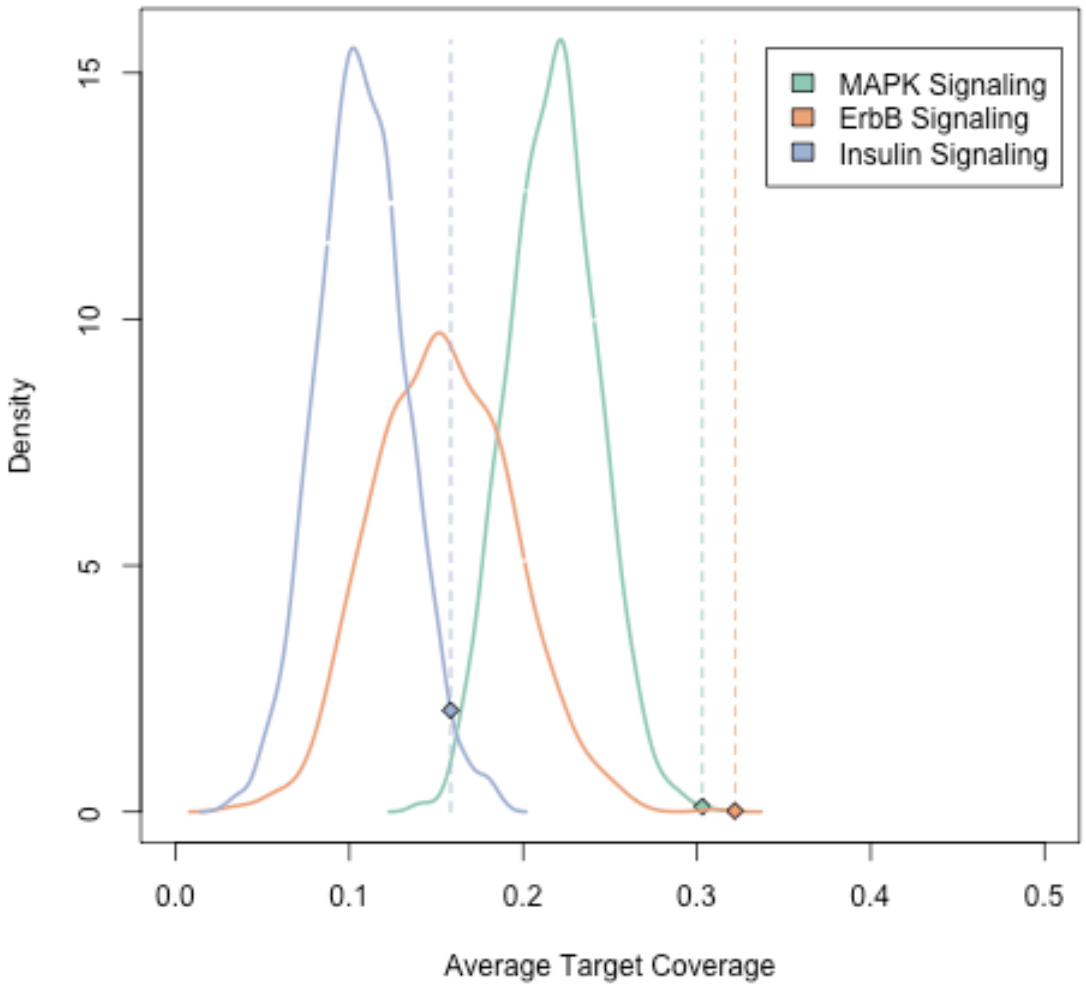

Figure 19 - Target Coverage in MAPK, ErbB and Insulin Signaling Pathways

MAPK, ErbB and Insulin Signaling pathways had a highly significant intra-pathway over-representation of intronic miRNAs targets. The figure shows the smoothed target coverage distribution obtained from random sampling, the dashed line indicates the actually observed target coverage (prediction agreement of at least two methods was required).

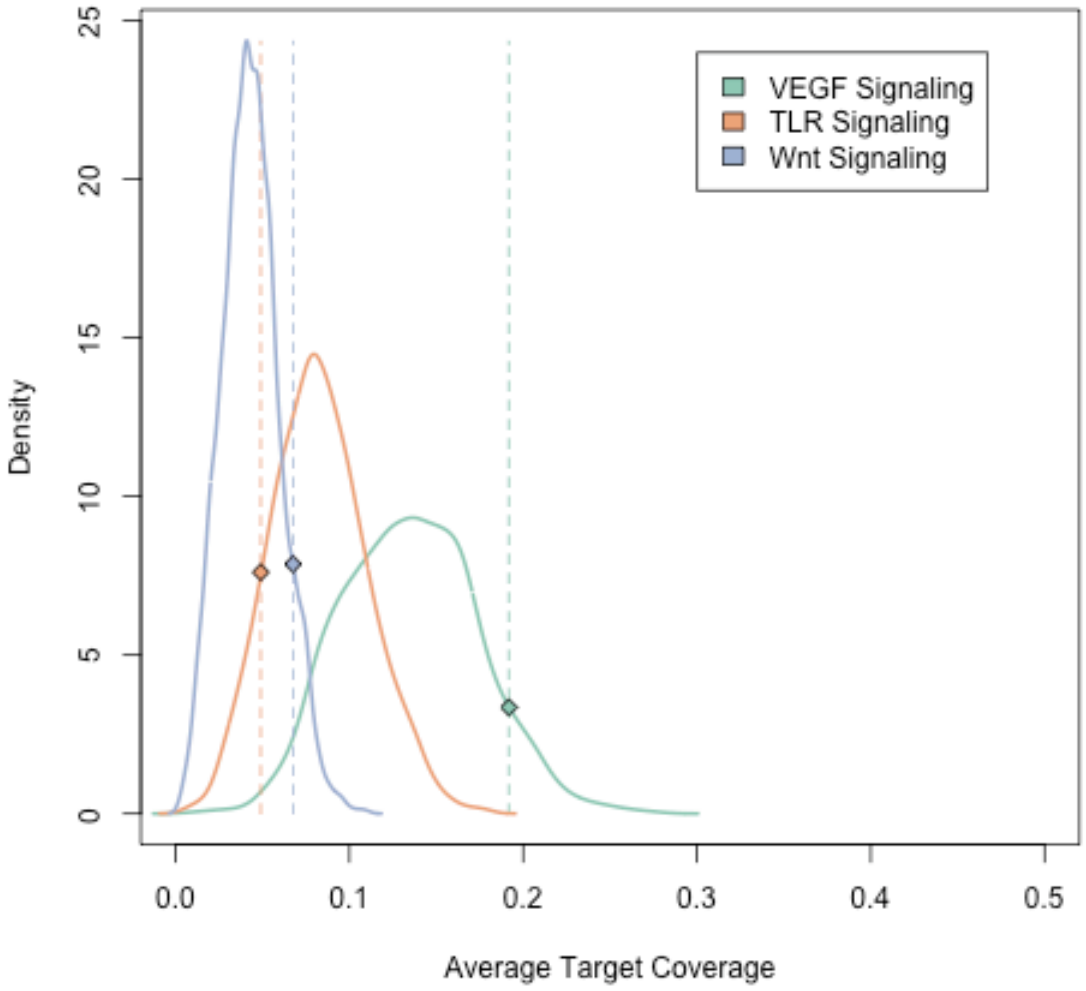

Figure 20 - Target Coverage in T-Cell, Jak-STAT, VEGF and Toll-Like-Receptor Signaling Pathways

VEGF, Toll-like Receptor (TLR) and Wnt signaling pathways were not found to be significant. Solid curves represent the distribution of the by chance expected target coverage, whereas the dashed lines show the observed target coverage.

4.2.4 Host – Target Correlation Suggests Role in Cancer Development

Blenkiron and coworkers [95] suggested that miRNA processing might be disturbed in cancer. They were able to show that some important enzymes involved in miRNA biogenesis were differentially expressed between tumor and normal samples, which might explain lower than expected correlation between host and miRNA expression levels [95].

In the setting of a negative feedback mechanism, this could have great impact, as the inhibitory control of the host would be attenuated. Integration of major KEGG pathway information with expression data from two publicly available datasets [102, 103] helped us investigate this issue. Assuming multi-order negative feedback (i.e. the intragenic miRNA does not target its own host, but functionally associated proteins), a host's expression levels and its miRNA's targets' expression levels would be negatively correlated, if the host and its miRNA were co-expressed. Correlation should be less pronounced or even positive, however, in tumor tissue, given reduced co-expression of host and miRNA gene.

KEGG ID "05215 – Prostate Cancer" contains a single known intronic miRNA host (*AKT2*), which is not predicted to be targeted by its intronic miRNA (*hsa-miR-641*). The correlations between host and predicted targets involved in and relevant to the pathway were calculated. Figure 21 shows a simplified representation based on the KEGG pathway information. Host and corresponding targets are color-coded, where the green oval indicates the host, *AKT2*, and yellow, orange, and red indicate whether two, three or four methods agreed on the target prediction. In line with the hypothesis of a negative feedback circuitry, targets of *hsa-miR-641* are to a great extent in close proximity to, and functional synergy with its host. A similar target pattern is exposed by both miRNAs, *hsa-miR-641* and *hsa-mir-634*, in the non small cell lung cancer pathway (Figure 22).

The correlation between host and target expression levels is shown in a two-bar plot. The first bar, labeled "N", represents the correlation between host and target in normal tissue. Similarly, the second bar, labeled "T", represents the correlation between host and target in tumorous tissue. In the prostate cancer dataset, seven of the fifteen targets are more negatively correlated in healthy tissue than in cancer. In three cases (*AKT3*, *AR*, and *CTNNB1*), one can observe a significant negative correlation in normal tissue, which is either non-significant or significantly positive in cancer. A similar pattern can be observed in the non small cell lung cancer pathway (Figure 22).

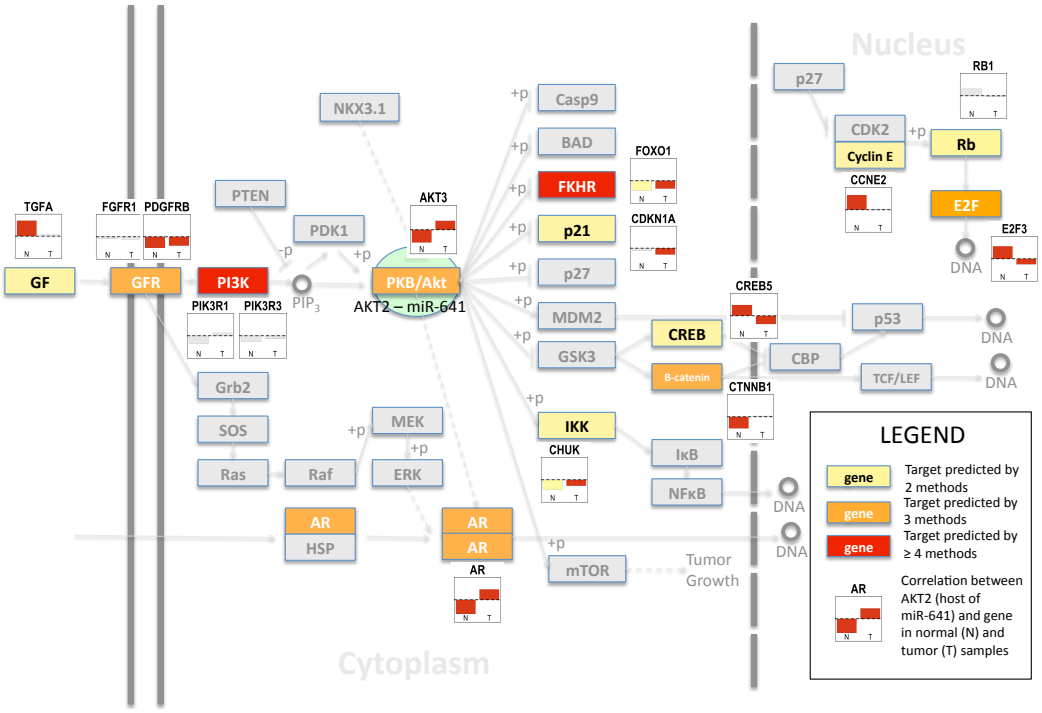

Figure 21 - Correlation of Host and Target in the Prostate Cancer Pathway

The PIK3/AKT2 signaling pathway plays a central role in prostate cancer. The majority of predicted targets of *hsa-miR-641* within the pathway appear to be in close proximity to, and functional synergy with its host, *AKT2*. Multiple potential targets show strong negative correlation with *AKT2* in normal tissue but weaker negative correlation, no correlation, or even positive correlation in tumor tissue (red bar: $p < 0.05$; yellow bar: $p < 0.10$; grey bar: $p \geq 0.10$).

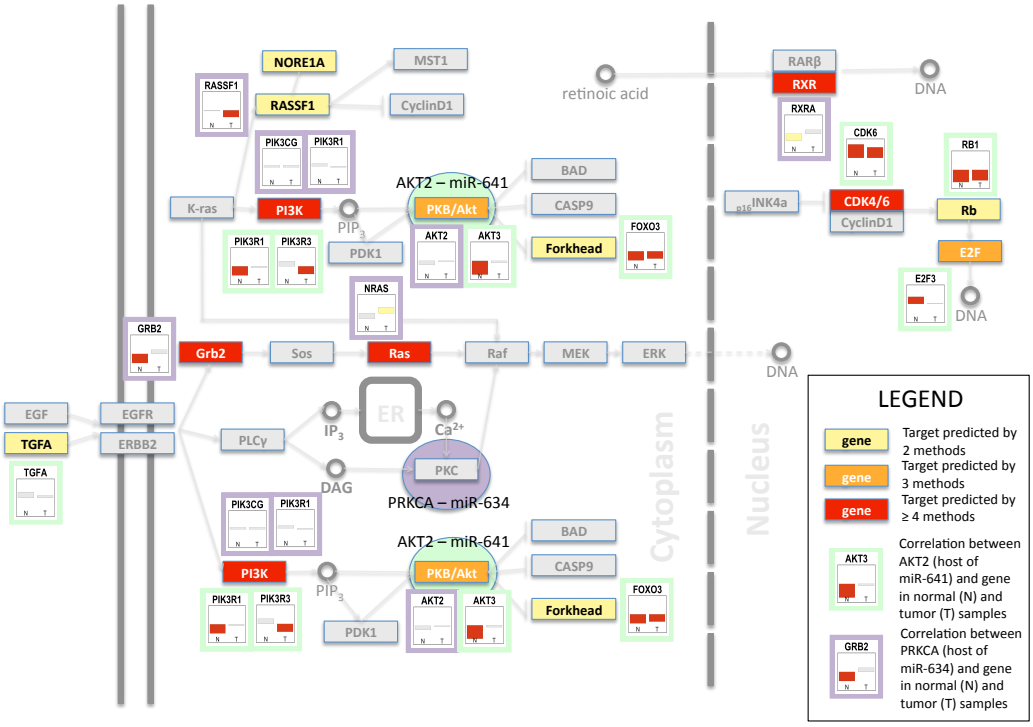

Figure 22 - Correlation of Host and Target in the Non Small Cell Lung Cancer Pathway

Similarly to what was found in the prostate cancer pathway, the majority of potential targets of *hsa-miR-641* and *hsa-miR-634* are close and functionally synergistic to their host genes. Several hypothesized targets are more strongly negatively correlated in normal tissue than in cancer (red bar: $p < 0.05$; yellow bar: $p < 0.10$; grey bar: $p \geq 0.10$).

5 Discussion

Since the first discovery of miRNAs, our understanding of biogenesis, target interaction and regulation has exponentially grown. In the recent past, it has been estimated that miRNAs that reside in intronic or exonic regions of other genes may be the dominating class [19]. However, functional aspects of intragenic miRNAs are still largely unknown.

5.1 Co-regulation Properties of Intronic miRNAs and Hosts

Little is known about the properties of co-regulation of intragenic miRNAs and their host genes. It is generally believed that both genes, host and miRNA, share regulatory control [7, 91]. However, recent reports accumulated evidence of post-transcriptional miRNA regulation [97, 98, 129], which raises uncertainty about biological means of co-transcription. After mapping miRNAs to known genes in RefSeq, we found that most intronic miRNAs are preferentially oriented in the same direction as their host gene (Table 4), significantly more than would be expected by chance. We showed that intronic miRNAs are not evenly distributed across the introns of their host genes, but have a positional bias towards the 5' introns. From a functional perspective, this finding integrates well with the idea that proximity to the start site of transcription may guarantee more stable transcription. Additionally, it is believed that the 5' introns of a gene contain regulatory elements [92], which supports our findings, considering that miRNAs themselves can be viewed as regulatory elements, albeit of a different kind. We also looked at expression correlation between miRNAs and their hosts in prostate cancer [101], where a significant averaged correlation between intragenic miRNAs and their hosts was observed. However, only 20% of [miRNA – host] pairs were individually significantly correlated (as opposed to 67% reported by Baskerville and Bartel in normal tissue [7]), consistent with previous reports that suggested altered post-transcriptional regulation at various levels in cancer tissue [95, 96, 98]. Even though the total number of 42 [miRNA – host] pairs is relatively low, we observed that significant correlation of expression levels was only observed in introns shorter than 8kb. This supports the idea

that some intronic miRNAs may be independently regulated when the intron is large enough to contain additional regulatory regions.

5.2 Functional Significance of Co-regulation

Co-regulation of intragenic miRNA and host through co-expression can be meaningfully explained in the context of either functional synergy or antagonism. In order to characterize the relationship between intronic miRNAs and their hosts, it is necessary to gather comprehensive information on functional aspects of host genes themselves, including their regulation, as well as functional aspects of miRNAs, which may be indirectly assessed by analysis of their targets.

Current knowledge about functional aspects of genes and their products is stored in biomedical ontologies, allowing the investigation of the association of a list of genes of interest to biological processes, as according to molecular function, localization within the cell, or biochemical pathways. We used Gene Ontology's ontologies "molecular function", "biological process", and "cellular compartment" [113] to first investigate the role of the hosts with respect to their function within the cell. We found an association with metabolic, biosynthetic, and gene regulative processes, as well as associations with cell compartments including synapsis, cell adherens and junctions, myofibrils and cytoskeleton. These categories capture major functional aspects of miRNAs in general, as is reflected by miRNA involvement in diseases such as cancer [130], muscle disorders [60], or neurodegenerative diseases [65]. The impact of miRNA in cell processes, via their host genes, was studied to further understand their functional role. Additionally, surveying KEGG biochemical pathways revealed that hosts of intronic miRNAs were associated with many signaling pathways, some of which are known to be involved in cancer.

Direct assessment of miRNA targets is difficult, as high throughput methods to comprehensively identify and validate targets for given miRNAs are still in development. However, some low to medium throughput experiments help us interpret our findings. Some researchers systematically over-expressed individual miRNAs and assessed changes in mRNA expression levels to infer miRNA-target interactions [36]. As described earlier,

however, all miRNA targets that will be translationally repressed cannot be captured by this method (false negatives). Also, miRNAs that target transcription factors for certain mRNAs may lead to false positives. Some authors developed theories about binding properties and tested these in several single miRNA target interaction experiments [84]. Conservation of potential mRNA binding sites across species has also been used to identify targets [77], but this approach misses those targets that are specific to a species. The knowledge of these experiments and hypotheses has been utilized in a variety of target prediction algorithms. To acquire a comprehensive list of potential target interactions, we combined predictions derived from these algorithms. Using the unified set of predictions, we could show that intragenic miRNAs tend to target their own hosts, supporting the concept of a first-order negative feedback regulation.

Even though our knowledge about current biochemical pathways and molecule interactions is still far from complete, we observed that intronic miRNAs seem to preferentially target molecules involved in the same biochemical pathway as their hosts, consistent with functional antagonism. A visual representation of the targets of *AKT2*'s intronic miRNA *hsa-miR-641*, for example, shows how components of many protein complexes involved in the signal transduction of growth factor signaling are potential targets of *hsa-miR-641* (Figure 21).

5.3 Potential Model of Cancer Development

Cancer encompasses a set of diseases that are characterized by uncontrolled growth of cells that are able to invade surrounding tissue and, by using lymphatic or blood vessels, metastasize to distinct parts of the body. In order to achieve this, these cells must be able to modify signals from surrounding cells or tissue and also signal transduction processes. Due to their regulatory function, miRNAs have been shown to be among the major players in cancer development [130]. In a recent study, Tavazoie et al. analyzed six miRNAs that were significantly under-expressed in breast cancer LM2 cells, as compared to normal breast tissue. Four of these miRNAs are intragenic [46]. The authors reported that loss of the intronic miRNA *hsa-miR-335*, which resides in intron 2 of its host gene *MEST*, lead to increased migration and invasion rates and hence increased metastatic

capacity. Additionally, they could show that *hsa-miR-126* (intron 7, host *EGFL7*) significantly reduced proliferation of breast cancer cells.

Some authors suggested dysregulation of miRNA biosynthesis in malignantly transformed cells [95, 96], leading to reduced correlation of expression levels with their hosts and hence explaining the apparently contradictory findings of Baskerville and Bartel [7] and Blenkiron et al [95]. If the assumption holds that intragenic miRNAs are functionally antagonistic to their hosts and that changes in miRNA biosynthesis such as those found in cancer reduce correlation of expression levels of miRNA and host, then one would expect to see a negative correlation between expression levels of host and target genes in normal tissue and a less negative or even positive correlation in cancerous tissue. This phenomenon was observed in two distinct datasets in different malignancies (Figure 21, Figure 22). A key to pathogenesis of both entities is the phosphatidylinositol 3-kinase(PIK3)/AKT signaling pathway, deregulation of which has been reported in several cancers, including prostate cancer [131], lung cancer [132], ovarian cancer [133, 134], breast cancer [134, 135] and colon tumors [133]. Modern drug therapy successfully targets *AKT* and *PI3K*. Whereas Noske et al. discovered that silencing *AKT2* through RNA interference leads to reduction in ovarian cancer cell proliferation [136], Maroulakou and coworkers reported accelerated development of polyoma middle T and ErbB2/Neu-driven mammary adenocarcinomas in mice after *AKT2* ablation [137]. Though these findings could appear to be contradictory at first, they could be explained by our model of an intronic miRNA-driven negative regulatory loop that is disinhibited in cancer. Whereas in the first experiment *AKT2* was targeted on mRNA level (and therefore mimicking the role of the corresponding intronic miRNA), the second experiment would downregulate both host mRNA and miRNA (if it exists in mouse), and would therefore disable negative feedback regulation by *hsa-miR-641*.

One should remember though that regulatory networks are far more complex in reality than what we are currently able to model. Transcription factors, enhancers, silencers, and epigenetic modifications play major roles in cancer development and may influence correlation among expression levels of host and target. Also, target prediction methods

are just predictions, and at this point we can only speculate about the true nature of events. Limitations of our model may justify, for example, why *Cyclin E* and *E2F* in Figure 21 show opposite behavior of what we would expect. For example, *Cyclin E* and *E2F* might actually not be targets of *hsa-miR-641*; there may also exist a stronger regulating element that controls expression levels, or the primary mode of silencing in that specific situation may be through translational repression. Nevertheless, it is interesting how key molecules in two different datasets display predicted correlation patterns and how our model can explain some recent findings in cancer research. Future steps include biochemical validation of the model. A starting point could be to show how selective inhibition and restoration of *hsa-miR-641* significantly modulates cell growth, proliferation, and survival as predicted by the model.

6 Summary

Even though intronic miRNAs have long been known, so far there has been no conclusive study determining the relationship between intronic miRNAs and their host genes and possible implications. The results reported here provide evidence that co-regulation through co-expression may be a key mechanism for at least a subset of intronic miRNAs to act as part of a negative feedback loop. When this mechanism is disrupted, abnormal cell development occurs, as is the case in cancer.

We show in this work, how computational analyses that integrate a variety of data and knowledge bases can be useful in the formulation of models that advance our understanding of disease processes. The fast pace by which technology to measure biological processes at a large scale is being developed, coupled with new informatics approaches that allow integrated analysis of large amounts of biological and clinical data is transforming the way biomedical experiments are being conducted, which is likely to accelerate the translation of scientific findings into critical advances in health care. The role of miRNAs in disease processes is just beginning to be understood, and much remains to be learned. This work represents a small, but important contribution towards elucidating the role of miRNAs in health and disease.

7 References

1. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.
2. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.
3. Pasquinelli, A.E., et al., *Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA.* Nature, 2000. **408**(6808): p. 86-9.
4. Lau, N.C., et al., *An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans.* Science, 2001. **294**(5543): p. 858-62.
5. Lee, R.C. and V. Ambros, *An extensive class of small RNAs in Caenorhabditis elegans.* Science, 2001. **294**(5543): p. 862-4.
6. Lagos-Quintana, M., et al., *Identification of novel genes coding for small expressed RNAs.* Science, 2001. **294**(5543): p. 853-8.
7. Baskerville, S. and D.P. Bartel, *Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes.* RNA, 2005. **11**(3): p. 241-7.
8. Lee, Y., et al., *The nuclear RNase III Drosha initiates microRNA processing.* Nature, 2003. **425**(6956): p. 415-9.
9. Lee, Y., et al., *MicroRNA genes are transcribed by RNA polymerase II.* EMBO J, 2004. **23**(20): p. 4051-60.
10. Borchert, G.M., W. Lanier, and B.L. Davidson, *RNA polymerase III transcribes human microRNAs.* Nat Struct Mol Biol, 2006. **13**(12): p. 1097-101.
11. Denli, A.M., et al., *Processing of primary microRNAs by the Microprocessor complex.* Nature, 2004. **432**(7014): p. 231-5.
12. Gregory, R.I., et al., *The Microprocessor complex mediates the genesis of microRNAs.* Nature, 2004. **432**(7014): p. 235-40.
13. Han, J., et al., *The Drosha-DGCR8 complex in primary microRNA processing.* Genes Dev, 2004. **18**(24): p. 3016-27.
14. Han, J., et al., *Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex.* Cell, 2006. **125**(5): p. 887-901.
15. Yi, R., et al., *Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs.* Genes Dev, 2003. **17**(24): p. 3011-6.
16. Bohnsack, M.T., K. Czaplinski, and D. Gorlich, *Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs.* RNA, 2004. **10**(2): p. 185-91.
17. Lund, E., et al., *Nuclear export of microRNA precursors.* Science, 2004. **303**(5654): p. 95-8.
18. Filipowicz, W. and V. Pogacić, *Biogenesis of small nucleolar ribonucleoproteins.* Curr Opin Cell Biol, 2002. **14**(3): p. 319-27.
19. Kim, Y. and V. Kim, *Processing of intronic microRNAs.* EMBO J, 2007. **26**(3): p. 775-783.
20. Bernstein, E., et al., *Role for a bidentate ribonuclease in the initiation step of RNA interference.* Nature, 2001. **409**(6818): p. 363-6.

21. Hutvágner, G., et al., *A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA.* Science, 2001. **293**(5531): p. 834-8.
22. Grishok, A., et al., *Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing.* Cell, 2001. **106**(1): p. 23-34.
23. Ketting, R.F., et al., *Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans.* Genes Dev, 2001. **15**(20): p. 2654-9.
24. Knight, S.W. and B.L. Bass, *A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans.* Science, 2001. **293**(5538): p. 2269-71.
25. Khvorova, A., A. Reynolds, and S.D. Jayasena, *Functional siRNAs and miRNAs exhibit strand bias.* Cell, 2003. **115**(2): p. 209-16.
26. Schwarz, D.S., et al., *Asymmetry in the assembly of the RNAi enzyme complex.* Cell, 2003. **115**(2): p. 199-208.
27. Nielsen, C., et al., *Determinants of targeting by endogenous and exogenous microRNAs and siRNAs.* RNA, 2007. **13**(11): p. 1894-1910.
28. Eulalio, A., E. Huntzinger, and E. Izaurralde, *Getting to the root of miRNA-mediated gene silencing.* Cell, 2008. **132**(1): p. 9-14.
29. Seggerson, K., L. Tang, and E.G. Moss, *Two genetic circuits repress the Caenorhabditis elegans heterochronic gene lin-28 after translation initiation.* Dev Biol, 2002. **243**(2): p. 215-25.
30. Maroney, P.A., et al., *Evidence that microRNAs are associated with translating messenger RNAs in human cells.* Nat Struct Mol Biol, 2006. **13**(12): p. 1102-7.
31. Nottrott, S., M.J. Simard, and J.D. Richter, *Human let-7a miRNA blocks protein production on actively translating polyribosomes.* Nat Struct Mol Biol, 2006. **13**(12): p. 1108-14.
32. Petersen, C.P., et al., *Short RNAs repress translation after initiation in mammalian cells.* Mol Cell, 2006. **21**(4): p. 533-42.
33. Kiriakidou, M., et al., *An mRNA m7G cap binding-like motif within human Ago2 represses translation.* Cell, 2007. **129**(6): p. 1141-51.
34. Chendrimada, T.P., et al., *MicroRNA silencing through RISC recruitment of eIF6.* Nature, 2007. **447**(7146): p. 823-8.
35. Wang, X. and I.M. El Naqa, *Prediction of both conserved and nonconserved microRNA targets in animals.* Bioinformatics, 2008. **24**(3): p. 325-32.
36. Linsley, P.S., et al., *Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression.* Mol Cell Biol, 2007. **27**(6): p. 2240-52.
37. Behm-Ansmant, I., et al., *mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes.* Genes Dev, 2006. **20**(14): p. 1885-98.
38. Giraldez, A.J., et al., *Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs.* Science, 2006. **312**(5770): p. 75-9.
39. Wu, L., J. Fan, and J.G. Belasco, *MicroRNAs direct rapid deadenylation of mRNA.* Proc Natl Acad Sci USA, 2006. **103**(11): p. 4034-9.
40. Eulalio, A., et al., *Target-specific requirements for enhancers of decapping in miRNA-mediated gene silencing.* Genes Dev, 2007. **21**(20): p. 2558-70.
41. Ma, L., J. Teruya-Feldstein, and R.A. Weinberg, *Tumour invasion and metastasis initiated by microRNA-10b in breast cancer.* Nature, 2007. **449**(7163): p. 682-8.

42. Schetter, A., et al., *MicroRNA Expression Profiles Associated With Prognosis and Therapeutic Outcome in Colon Adenocarcinoma.* JAMA: The Journal of the American Medical Association, 2008. **299**(4): p. 425-436.
43. Yan, L.X., et al., *MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis.* RNA, 2008. **14**(11): p. 2348-60.
44. Kluiver, J., et al., *BIC and miR-155 are highly expressed in Hodgkin, primary mediastinal and diffuse large B cell lymphomas.* J Pathol, 2005. **207**(2): p. 243-9.
45. Gironella, M., et al., *Tumor protein 53-induced nuclear protein 1 expression is repressed by miR-155, and its restoration inhibits pancreatic tumor development.* Proc Natl Acad Sci USA, 2007. **104**(41): p. 16170-5.
46. Tavazoie, S.F., et al., *Endogenous human microRNAs that suppress breast cancer metastasis.* Nature, 2008. **451**(7175): p. 147-52.
47. Shi, L., et al., *hsa-mir-181a and hsa-mir-181b function as tumor suppressors in human glioma cells.* Brain Res, 2008.
48. Tazawa, H., et al., *Tumor-suppressive miR-34a induces senescence-like growth arrest through modulation of the E2F pathway in human colon cancer cells.* Proc Natl Acad Sci USA, 2007. **104**(39): p. 15472-7.
49. Cui, C., et al., *Prediction and identification of herpes simplex virus 1-encoded microRNAs.* J Virol, 2006. **80**(11): p. 5499-508.
50. Pfeffer, S., et al., *Identification of microRNAs of the herpesvirus family.* Nat Meth, 2005. **2**(4): p. 269-76.
51. Pfeffer, S., et al., *Identification of virus-encoded microRNAs.* Science, 2004. **304**(5671): p. 734-6.
52. Xing, L. and E. Kieff, *Epstein-Barr virus BHRF1 micro- and stable RNAs during latency III and after induction of replication.* J Virol, 2007. **81**(18): p. 9967-75.
53. Omoto, S., et al., *HIV-1 nef suppression by virally encoded microRNA.* Retrovirology, 2004. **1**: p. 44.
54. Hariharan, M., et al., *Targets for human encoded microRNAs in HIV genes.* Biochemical and Biophysical Research Communications, 2005. **337**(4): p. 1214-8.
55. Scaria, V., et al., *Host-virus interaction: a new role for microRNAs.* Retrovirology, 2006. **3**: p. 68.
56. Huang, J., et al., *Cellular microRNAs contribute to HIV-1 latency in resting primary CD4+ T lymphocytes.* Nat Med, 2007. **13**(10): p. 1241-7.
57. Lecellier, C.H., et al., *A cellular microRNA mediates antiviral defense in human cells.* Science, 2005. **308**(5721): p. 557-60.
58. Watanabe, Y., et al., *Computational analysis of microRNA-mediated antiviral defense in humans.* FEBS Letters, 2007.
59. McCarthy, J.J., *MicroRNA-206: The skeletal muscle-specific myomiR.* Biochim Biophys Acta, 2008. **1779**(11): p. 682-91.
60. Eisenberg, I., et al., *Distinctive patterns of microRNA expression in primary muscular disorders.* Proc Natl Acad Sci USA, 2007. **104**(43): p. 17016-21.
61. Carè, A., et al., *MicroRNA-133 controls cardiac hypertrophy.* Nat Med, 2007. **13**(5): p. 613-8.
62. Divakaran, V. and D.L. Mann, *The emerging role of microRNAs in cardiac remodeling and heart failure.* Circ Res, 2008. **103**(10): p. 1072-83.

63. Fiore, R., G. Siegel, and G. Schratt, *MicroRNA function in neuronal development, plasticity and disease.* Biochim Biophys Acta, 2008. **1779**(8): p. 471-8.
64. Hansen, T., et al., *Brain Expressed microRNAs Implicated in Schizophrenia Etiology.* PLoS ONE, 2007.
65. Niwa, R., et al., *The expression of the Alzheimer's amyloid precursor protein-like gene is regulated by developmental timing microRNAs and their targets in Caenorhabditis elegans.* Dev Biol, 2008. **315**(2): p. 418-25.
66. Wang, G., et al., *Variation in the miRNA-433 binding site of FGF20 confers risk for Parkinson disease by overexpression of alpha-synuclein.* Am J Hum Genet, 2008. **82**(2): p. 283-9.
67. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.
68. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
69. Ritchie, M.E., et al., *A comparison of background correction methods for two-colour microarrays.* Bioinformatics, 2007. **23**(20): p. 2700-7.
70. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.
71. Schadt, E.E., et al., *Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.* J Cell Biochem Suppl, 2001. **Suppl 37**: p. 120-5.
72. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics (Oxford, England), 2003. **4**(2): p. 249-64.
73. Wolfinger, R.D., et al., *Assessing gene significance from cDNA microarray expression data via mixed models.* J Comput Biol, 2001. **8**(6): p. 625-37.
74. Kerr, M.K. and G.A. Churchill, *Experimental design for gene expression microarrays.* Biostatistics (Oxford, England), 2001. **2**(2): p. 183-201.
75. Churchill, G.A., *Fundamentals of experimental design for cDNA microarrays.* Nat Genet, 2002. **32 Suppl**: p. 490-5.
76. Churchill, G.A. and B. Oliver, *Sex, flies and microarrays.* Nat Genet, 2001. **29**(4): p. 355-6.
77. Lewis, B.P., et al., *Prediction of mammalian microRNA targets.* Cell, 2003. **115**(7): p. 787-98.
78. Hofacker, I.L., et al., *Fast folding and comparison of RNA secondary structures.* Monatshefte für Chemie/Chemical Monthly, 1994.
79. Enright, A.J., et al., *MicroRNA targets in Drosophila.* Genome Biol, 2003. **5**(1): p. R1.
80. John, B., et al., *Human MicroRNA targets.* PLoS Biol, 2004. **2**(11): p. e363.
81. 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.
82. Rigoutsos, I. and A. Floratos, *Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm.* Bioinformatics, 1998. **14**(1): p. 55-67.
83. Doolittle, R.F., et al., *Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor.* Science, 1983. **221**(4607): p. 275-7.
84. Kertesz, M., et al., *The role of site accessibility in microRNA target recognition.* Nat Genet, 2007. **39**(10): p. 1278-84.

85. Wang, X., *miRDB: a microRNA target prediction and functional annotation database with a wiki interface.* RNA, 2008. **14**(6): p. 1012-7.
86. Krek, A., et al., *Combinatorial microRNA target predictions.* Nat Genet, 2005. **37**(5): p. 495-500.
87. Rajewsky, N., *microRNA target predictions in animals.* Nat Genet, 2006. **38 Suppl**: p. S8-13.
88. Schroeder, M.D., et al., *Transcriptional control in the segmentation gene network of Drosophila.* PLoS Biol, 2004. **2**(9): p. E271.
89. Sethupathy, P., B. Corda, and A. Hatzigeorgiou, *TarBase: A comprehensive database of experimentally supported animal microRNA targets.* RNA, 2006. **12**(2): p. 192-7.
90. Nott, A., S.H. Meislin, and M.J. Moore, *A quantitative analysis of intron effects on mammalian gene expression.* RNA, 2003. **9**(5): p. 607-17.
91. Rodriguez, A., et al., *Identification of mammalian microRNA host genes and transcription units.* Genome Research, 2004. **14**(10A): p. 1902-10.
92. Xie, X., et al., *Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.* Nature, 2005. **434**(7031): p. 338-45.
93. Li, S., P. Tang, and W. Lin, *Intronic microRNA: discovery and biological implications.* DNA and Cell Biology, 2007. **26**(4): p. 195-207.
94. Saini, H.K., S. Griffiths-Jones, and A.J. Enright, *Genomic analysis of human microRNA transcripts.* Proc Natl Acad Sci USA, 2007. **104**(45): p. 17719-24.
95. Blenkiron, C., et al., *MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype.* Genome Biol, 2007. **8**(10): p. R214.
96. Muralidhar, B., et al., *Global microRNA profiles in cervical squamous cell carcinoma depend on Drosha expression levels.* J Pathol, 2007. **212**(4): p. 368-77.
97. Obernosterer, G., et al., *Post-transcriptional regulation of microRNA expression.* RNA, 2006. **12**(7): p. 1161-7.
98. Thomson, J.M., et al., *Extensive post-transcriptional regulation of microRNAs and its implications for cancer.* Genes Dev, 2006. **20**(16): p. 2202-7.
99. Barik, S., *An intronic microRNA silences genes that are functionally antagonistic to its host gene.* Nucleic Acids Research, 2008. **36**(16): p. 5232-41.
100. Megraw, M., et al., *MicroRNA promoter element discovery in Arabidopsis.* RNA, 2006. **12**(9): p. 1612-9.
101. Prueitt, R., et al., *Expression of microRNAs and protein-coding genes associated with perineural invasion in prostate cancer.* Prostate, 2008. **68**(11): p. 1152-1164.
102. Wallace, T.A., et al., *Tumor immunobiological differences in prostate cancer between African-American and European-American men.* Cancer Research, 2008. **68**(3): p. 927-36.
103. Su, L.J., et al., *Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme.* BMC Genomics, 2007. **8**: p. 140.
104. Rao, Y., et al., *A comparison of normalization techniques for microRNA microarray data.* Statistical applications in genetics and molecular biology, 2008. **7**: p. Article22.
105. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Research, 2003. **31**(4): p. e15.
106. McGee, M. and Z. Chen, *Parameter estimation for the exponential-normal convolution model for background correction of affymetrix GeneChip data.* Statistical applications in genetics and molecular biology, 2006. **5**: p. Article24.

107. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Research, 2005. **33**(Database issue): p. D501-4.
108. Griffiths-Jones, S., *The microRNA Registry.* Nucleic Acids Research, 2004. **32**(Database issue): p. D109-11.
109. Griffiths-Jones, S., *miRBase: the microRNA sequence database.* Methods Mol Biol, 2006. **342**: p. 129-38.
110. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Research, 2006. **34**(Database issue): p. D140-4.
111. Griffiths-Jones, S., et al., *miRBase: tools for microRNA genomics.* Nucleic Acids Research, 2008. **36**(Database issue): p. D154-8.
112. Ambros, V., et al., *A uniform system for microRNA annotation.* RNA, 2003. **9**(3): p. 277-9.
113. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
114. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Research, 2003. **13**(11): p. 2498-504.
115. Maere, S., K. Heymans, and M. Kuiper, *BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.* Bioinformatics, 2005. **21**(16): p. 3448-9.
116. R Development Core Team (2008), *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
117. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol, 2004. **5**(10): p. R80.
118. Zhang, J., V. Carey, and R. Gentleman, *An extensible application for assembling annotation for genomic data.* Bioinformatics, 2003.
119. Falcon, S. and R. Gentleman, *Using GOstats to test gene lists for GO term association.* Bioinformatics, 2007. **23**(2): p. 257-8.
120. Consortium, E.P., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
121. Luo, Y. and S. Zhang, *Computational prediction of amphioxus microRNA genes and their targets.* Gene, 2008.
122. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Research, 2000. **28**(1): p. 27-30.
123. Kanehisa, M., et al., *KEGG for linking genomes to life and the environment.* Nucleic Acids Research, 2008. **36**(Database issue): p. D480-4.
124. Kanehisa, M., et al., *From genomics to chemical genomics: new developments in KEGG.* Nucleic Acids Research, 2006. **34**(Database issue): p. D354-7.
125. Keyse, S.M., *Dual-specificity MAP kinase phosphatases (MKPs) and cancer.* Cancer Metastasis Rev, 2008. **27**(2): p. 253-61.
126. Fry, W.H., et al., *Mechanisms of ErbB receptor negative regulation and relevance in cancer.* Exp Cell Res, 2008.
127. Roderick, H.L. and S.J. Cook, *Ca2+ signalling checkpoints in cancer: remodelling Ca2+ for cancer cell proliferation and survival.* Nature Reviews Cancer, 2008. **8**(5): p. 361-75.
128. Strimpakos, A.S., et al., *The role of mTOR in the management of solid tumors: An overview.* Cancer Treat Rev, 2008.

129. Lee, E.J., et al., *Systematic evaluation of microRNA processing patterns in tissues, cell lines, and tumors.* RNA, 2008. **14**(1): p. 35-42.
130. Ma, L. and R.A. Weinberg, *MicroRNAs in malignant progression.* Cell Cycle, 2008. **7**(5): p. 570-2.
131. Boormans, J.L., et al., *An activating mutation in AKT1 in human prostate cancer.* Int J Cancer, 2008. **123**(11): p. 2725-6.
132. Forgacs, E., et al., *Mutation analysis of the PTEN/MMAC1 gene in lung cancer.* Oncogene, 1998. **17**(12): p. 1557-65.
133. Philp, A.J., et al., *The phosphatidylinositol 3'-kinase p85alpha gene is an oncogene in human ovarian and colon tumors.* Cancer Research, 2001. **61**(20): p. 7426-9.
134. Bellacosa, A., et al., *Molecular alterations of the AKT2 oncogene in ovarian and breast carcinomas.* Int J Cancer, 1995. **64**(4): p. 280-5.
135. Sun, M., et al., *Phosphatidylinositol-3-OH Kinase (PI3K)/AKT2, activated in breast cancer, regulates and is induced by estrogen receptor alpha (ERalpha) via interaction between ERalpha and PI3K.* Cancer Research, 2001. **61**(16): p. 5985-91.
136. Noske, A., et al., *Specific inhibition of AKT2 by RNA interference results in reduction of ovarian cancer cell proliferation: increased expression of AKT in advanced ovarian cancer.* Cancer Lett, 2007. **246**(1-2): p. 190-200.
137. Maroulakou, I.G., et al., *Akt1 ablation inhibits, whereas Akt2 ablation accelerates, the development of mammary adenocarcinomas in mouse mammary tumor virus (MMTV)-ErbB2/neu and MMTV-polyoma middle T transgenic mice.* Cancer Research, 2007. **67**(1): p. 167-77.