REGULATION OF GENE EXPRESSION BY MICRORNAS: TARGETING SPECIFICITY, KINETICS AND FUNCTION

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

JEAN ALBERT RENÉ HAUSSER

aus Frankreich

Basel, 2011

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Mihaela Zavolan, Prof. Sven Bergmann

Basel, den 14.12.2010

Prof. Dr. Martin Spiess
Dekan

Jean Albert René Hausser: *Regulation of gene expression by microRNAs: targeting specificity, kinetics and function,* PhD thesis, 2011

ABSTRACT

Understanding gene regulation is a central question of molecular biology. For decades, gene expression was thought to be controlled by a complex network of proteins called transcription factors. But ten years ago, microRNAs (miRNAs), a distinct class of short, evolutionarily-conserved non-coding RNAs were found to regulate gene expression. Hundreds of miRNAs have since then been discovered in species ranging from plants to nematodes to mammals, where they regulate diverse biological processes such as development, metabolism, immunity, cell cycle. MicroRNAs load into the Argonaute protein of the RNA-Induced Silencing Complex (RISC) and provide binding specificity to it. Upon guiding the RISC to a complementary motif in the 3′ untranslated transcribed region (UTR) of a mRNA, miRNAs inhibit the translation and increase the decay rate of the target mRNA.

While the molecular machinery required for miRNA action is well characterized, the biological function of the miRNAs identified so far remains unknown. Neither do we know through what target genes miRNAs achieve their biological function. The most common approach to this question consists in identifying genes that are differentially expressed following the experimental perturbation of the expression of a given miRNA by means of genetic knock-out or transfection. Perturbing the expression of a single miRNA has important side-effects on gene expression, but this problem can be partly addressed by crossing the genes responding to the miRNA perturbation with computational miRNA target predictions. In this thesis, we first illustrate how such a combined experimental and computational approach can be used to understand how the miR-375 miRNA controls glucose homeostasis.

However, in practice, extracting direct, functional miRNA targets from miRNA perturbation experiments and computational predictions is a difficult task because state-of-the-art computational predictions yield large amounts of false-positives. We therefore set to improve the accuracy of computational predictions by inferring what sequence and structure properties characterize functional miRNA binding sites in a large number of miRNA perturbation experiments. We then combined these properties into an algorithm that is most accurate at miRNA target prediction. Also, we show that miRNA binding sites carried by mRNAs that respond to miRNA perturbation share the same properties as miRNA binding sites that are under evolutionary selective pressure, suggesting that miRNA binding sites may have been shaped by evolution to favor mRNA degradation. Further analyses also lead to the view that the temporal aspects of miRNA regulation may be far more important to the miRNA target identification problem than previously thought, especially for experiments measuring the effects of miRNA perturbation at the protein level, where taking the temporal aspects of miRNA regulation into account appears necessary both during experimental design and subsequent data analysis.

While measurements from combined miRNA perturbation experiments and omics assays are crucial to determining what genes are regulated by a given miRNA, they are contaminated by side-effects and do not provide information on the precise location of the miRNA binding

site within the 3′ UTR of the target genes. To address these problems, we introduce PAR-CLIP, a combination of biochemical and computational methods to identify miRNA binding sites in high-throughput. The mRNA-miRNA-Argonaute ternary complex are first cross-linked. The ternary complex is then immuno-precipitated and the unprotected RNA eliminated by enzymatic digestion. Finally, ultra high-throughput sequencing of the remaining RNA and computational processing of the resulting sequencing libraries reveals the precise mRNA regions bound by miRNAs. PAR-CLIP does not require miRNA perturbation and makes it possible to identify thousands of miRNA binding sites in one experiment, with nucleotide resolution.

In summary, the present thesis establishes methods that make it possible to map miRNA-mRNA interactions with high accuracy in the spatial domain, and paves the way for future investigation of miRNA-mediated gene regulation in the temporal domain. These methods will be useful in understanding the miRNA-mRNA interactions underlying the implication of miRNAs in the regulation of biological processes.

ZUSAMMENFASSUNG

Die Regulation der Genexpression ist eine zentrale Frage der molekularen Biologie. Während Jahrzehnten wurde angenommen, dass die Expression der Genen von komplexen Netzwerken kontrolliert wird, die aus Proteinen, so genannten Transkriptionsfaktoren bestehen. Vor zehn Jahren wurde entdeckt, dass microRNAs (miRNAs) eine eigene Klasse kleiner, in der Evolution konservierter, nicht-codierender RNA bilden, die Genexpression regulieren. Seitdem wurden hunderte von miRNAs in Organismen, unter ihnen Pflanzen, Nematoden und Säugetieren entdeckt, wo sie diverse biologische Prozesse wie Entwicklung, Metabolismus, Immunität, Zellzyklus regulieren. MicroRNAs binden an die Argonaute Protein vom RNA-Induced Silencing Complex (RISC) und bestimmen so die Bindungsspezifität der Argonaute. MiRNAs führen dann den RISC zu einem komplementären Motif der 3′ untranslatierten Region (UTR) einer mRNA, was zur Inhibition der Translation und zur Erhöhung der Zerfallsrate der gebundenen mRNA führt.

Während die molekularen Mechanismen der Genexpressionsregulation durch miRNAs identifiziert wurden, bleibt die biologische Funktion einer grossen Mehrheit der miRNAs, die so weit entdeckt wurden, unbekannt. Es ist zudem unklar, durch welche Gene die miRNA ihre Funktion ausüben. Die häufigste Herangehensweise, diese Frage zu beantworten ist die Identifikation von Genen, deren Expression durch eine gegebene miRNA gestört wird. Genetische Knock-Outs oder Transfektionen sind experimentelle Mittel um die Expression zu stören. Die Expression einer einzelnen miRNAs zu stören kann erhebliche sekundäre Effekte auf die Expression von Genen haben. Durch die Kreuzung von miRNA abhängigen, differentiel exprimierten Genen mit rechnergeschützten miRNA Bindundungsstellenvorhersagen (rmBV) kann dieses Problem teilweise gelöst werden. In dieser Dissertation wurde diese Strategie eingesetzt um zu untersuchen, wie miRNA-375 die Glukosehomeostase kontrolliert.

In der Praxis ist es jedoch eine anspruchsvolle Arbeit, direkte, funktionelle miRNA Zielgene aus miRNA-Störungsexperimenten und rmBV zu extrahieren da rmBV in der Regel einen hohen Anteil an falsch

Positiven liefern. Wir verbesserten die Genauigkeit der rmBV indem wir die Sequenz- und Struktureigenschaften von funktionellen miRNA Bindungsstellen aus einer grossen Anzahl von miRNA Störungsexperimenten charakterisierten. Die identifizierten Eigenschaften wurden dann mit dem Algorithmus zur Vorhersage der miRNA-Bindungsstellen kombiniert, der bei der Identifikation von Ziel-miRNA am genauesten ist. Zudem zeigen wir, dass miRNA Bindungsstellen von miRNA-abhängigen mRNAs dieselben Eigenschaften aufweisen wie Bindungsstellen, welche unter evolutionärem Selektionsdruck stehen. Das führt zur Hypothese, dass miRNA Bindungsstellen durch die Evolution umgeformt wurden, um den mRNA Zerfall zu bevorzugen. Weitere Analysen führten zur Auffassung, dass die zeitlichen Aspekte der miRNA Regulation viel wichtiger sein könnten als bisher angenommen. Dies speziell für Experimente, die den Effekt der miRNA Störung auf der Ebene der Proteine messen. Bei diesen Experimenten scheint es unentbehrlich zu sein, während der Planung und Datenanalyse Rücksicht auf die zeitlichen Aspekte der miRNA Regulation zu nehmen.

Messungen aus kombinierten miRNA Störungsexperimenten und Omics-Versuchen sind ausschlaggebend um festzustellen welche Gene von einer bestimmten miRNA reguliert werden. Sie leiden jedoch darunter, dass sie von sekundären Effekten gestört werden und dass sie keine Information über die genaue Lokalisation der miRNA Bindungsstellen liefern. Um diese Probleme zu lösen wurde die PAR-CLIP Methode entwickelt. Dies ist eine Kombination aus biochemischen und rechnergestützten Methoden um miRNA Bindungsstellen in hohen Datendurchsätzen zu identifizieren. Die ternären mRNA-miRNA-Argonaute Komplexe werden erst kovalent gebunden, dann immuno-prezipitiert. Danach wird die ungeschützte RNA in einem enzymatischen Verdau eliminiert. Schlussendlich wird die verbleibende RNA sequenziert und durch rechnergestützte Verarbeitung der Sequenzierdaten wird festgestellt, welche spezifischen mRNA Regionen von miRNAs gebunden werden. PAR-CLIP benötigt keine miRNA Störung und ermöglicht die Identifizierung tausender miRNA Bindungsstellen Nukleotid-Auflösend in einem einzigen Versuch.

Zusammengefasst führt diese Dissertation Methoden ein, mit denen sich miRNA-mRNA Wechselwirkungen mit hoher räumlicher Genauigkeit kartografisieren lassen. Zudem öffnet sie den Weg für zukünftige Untersuchungen von zeitlichen Domänen in der miRNA vermittelten Genregulation. Diese Methoden werden entscheidend zum Verständnis der miRNA-mRNA Wechselwirkungen beitragen und den Einfluss der miRNA in der Regulation biologischer Prozesse betonen.

ACKNOWLEDGMENTS

I am much indebt to Lukas Burger, Lukasz Jaskiewicz, Dimos Gaidatzis, Phil Arnold, Piotr Balvierz, Mohsen Khorshid, Philip Berninger, Jose Ignacio Molina Clemente and Christoph Rodak for the scientific discussions, for sharing their knowledge, their views and their ideas on scientific questions and technical challenges. Their contribution was crucial to this work. Many thanks also to Tabitha Bucher for her warm support and for helping me with the German version of the abstract.

I am very grateful to Thomas Tuschl for inviting me to his lab for two intense and fascinating weeks of wet lab work, to Markus Landthaler for patiently supervising me during that time, and to Francesca Bersani for emergency late-night help with spectrometers, incubators and petri dishes.

Also, many thanks to Sven Bergmann and Markus Stoffel for their advice as part of my thesis committee.

Special thanks to Erik van Nimwegen for crucial contributions and out-of-the-box ideas to the present thesis. Talking to him about my scientific problems was always an enlightening, challenging and educative experience, which always ended with me leaving his office with enough new ideas to follow up on for months.

Warm thanks to Mihaela Zavolan, for her incredible availability, reactivity and dedication 24/7, her encouragements and support to travel and attend (on several occasions life changing) scientific events, her contagious energy, her smart and involved advising. These past four years were the most interesting time of my life, thanks for making this possible.

Last but not least, many thanks to my family and friends, for their warm support and encouragements during these four years.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

How gene expression is controlled in living cells has probably been the most central question of molecular biology in the last 50 years. For instance, the human body is made of 10^{12} human cells, all of which virtually share an identical genetic material, which is carried by the DNA and packaged in 23 chromosome pairs. Yet, these cells can be divided in cell types such as epithelial cells, neurons, myocytes, endocrine cells, immune cells (macrophages, lymphocytes, etc.), erythrocytes, which differ widely in morphology and function. The mechanism through which so much phenotypical diversity can be obtained from identical genetic material is gene expression control: different cell types express different genes, which determine the morphology and function of cells [2].

Gene expression control is also crucial in determining how cells react to a changing environment: a brutal depletion in a certain type of nutrient may require the production of an enzyme that makes it possible for the cell to metabolize an alternative type of nutrient. It is in this context that the first mechanism of gene expression control was characterized in bacteria [105, 58]. Many pathologies also have a deep connection with gene expression. For instance, viruses are parasites that replicate themselves by hijacking the gene expression machinery of infected cells, tricking the mechanisms of gene expression control of the host cell into expressing the viral genes. In cancer, cumulating accidental alterations to the genetic material can lead to defects in the expression of genes that are key to controlling the most basic cellular function such as growth, division and death, ultimately resulting in uncontrolled proliferation, migration and foreign tissue invasion [195]. Gene regulation control is known today to be — at least in part — the product of the action of proteins called transcription factors which bind the promoter region of genes to induce or repress the transcription of DNA by the RNA polymerase [28, 183]. Transcription factors bind specific DNA sequences [51] and their expression is itself regulated by transcription factors, which results in a genetic regulatory network whose function is to control gene expression in the cell.

The control of transcription was the first mechanism of gene expression control to be discovered. But it was not long until other mechanisms were proposed. Britten and Davidson [25] proposed a theory in which gene expression is controlled by intermolecular RNA-RNA pairing. A few years later, Heywood and Kennedy [100] found experimental evidence that a so-called "translation control RNA" (tcRNA) interacts with the myosin mRNA to inhibit its translation. This was the first experimental evidence of a non-coding RNA regulating gene expression at the post-transcriptional level. With a reported molecular weight of 10000, it is not clear whether the tcRNA of Heywood and Kennedy [100] may have been the first member of the large family of small, non-coding RNAs formed by microRNAs, whose molecular weight is more in the 6800 – 8000 range. In any case, the report did not gather a lot of attention. The role of RNAs in gene expression remained limited to that of passive, intermediate carriers of the genetic

information (messenger RNAs – mRNAs), or to essential enzymatic and co-factor functions in protein translation (transfer RNAs – tRNAs, ribosomal RNAs – rRNAs).

This view started shifting dramatically with the discovery of microRNAs (miRNAs), which are short, non-coding RNAs that repress gene expression at the post-transcriptional level. The first miRNA, lin-4, was found by genetic screens by [131, 231] in *Caenorhabditis elegans* and plays an important role in the development of the nematode by repressing the lin-14 heterochronic gene at the transition between the first and second larval stage. To date, hundreds of miRNAs have been discovered in a broad range of species ranging from plants to metazoans. A substantial fraction of them are conserved over long evolutionary distances [126]. Even some DNA viruses encode miRNAs [169], that act to regulate viral life cycle in the host cell as well as the expression of host genes [206, 159, 185, 79].

MiRNA biogenesis progresses through multiple steps and involves a collection of enzymes and transport proteins. MiRNAs are processed by the Drosha enzyme [132] from hairpin structures that occur in longer coding or non-coding transcripts. The pre-miRNA hairpins are exported out of the nucleus, further sliced into a double-stranded RNA by the Dicer enzyme and are loaded into the Argonaute protein of the RNA-Induced Silencing Complex (RISC) [84, 103, 157]. They confer target recognition specificity to the Argonaute protein and guide the RISC to miRNA recognition elements located mostly in the 3′ untranslated regions (3′ UTRs) of mRNAs. The binding of RISC to a mRNA results in a rapid repression of translation, decapping, deadenylation, and ultimate degradation of the target mRNA [66, 62]. With each miRNAs targeting the mRNAs of a specific set of genes, the view emerged that miRNAs form a new layer of gene regulatory networks on top of transcription control. As primary location for miRNA binding, 3′ UTRs of mRNAs are now considered to play the same role in post-transcriptional regulation as promoters do in transcription control.

Many fundamental biological processes, such as metabolism [170, 123], embryogenesis [78], cell cycle [140], cancer [30, 97], epigenetic modification [198, 50], and immunity [210, 162] are now known to be regulated by miRNAs. Given that the functions of many miRNAs that have been isolated in sequencing studies remain to be characterized, one can speculate that many more biological processes will be found to be under miRNA control. Therefore, characterizing miRNA expression and identification of miRNA targets is an important problem. A variety of platforms such as microarrays [31, 155], Sanger sequencing of small RNA libraries [126, 129] or next generation sequencing [89, 164] can be deployed to identify miRNAs and characterize their expression. These methods will not be discussed in this thesis. Here we rather focus on the question of understanding the function of miRNAs. As miRNAs are regulatory molecules that repress gene expression, understanding their function requires to understand which genes are repressed by individual miRNAs.

Many studies addressed the question of how miRNAs find their targets. While in plants miRNAs bind to nearly perfectly complementary targets [180], in metazoans it appears that most of the targeting specificity comes from the 7-8 nucleotides at the 5′ end of the miRNA, also known as the miRNA "seed" [134, 137, 24]. Because miRNAs are conserved in evolution and were originally discovered because of their

fundamental role in development, one can make the assumption that functional, physiologically relevant miRNA binding sites are under evolutionary selective pressure. Indeed, many miRNA target prediction methods make this assumption, aside from requiring extensive pairing of the miRNA seed, and/or unusually low free energy of binding between the miRNA and the mRNA [203, 134, 59, 176]. Under these constraints, the average number of predicted targets per miRNA is in the range of hundreds [135, 154]. This number is in striking contrast with the number of targets that have been so far validated for any individual miRNA, in part because experimental validation requires intense work.

Chapter 2 illustrates how the biological function and the regulatory mechanism of miRNAs can be elucidated in the context of the regulation of glucose homeostasis by the miR-375 miRNA. The study combines high-throughput measurements of the consequences of perturbing the expression of miR-375 on gene expression. Genes whose expression was altered in response to perturbing the expression of miR-375 were computationally screened for potential miRNA binding sites, which resulted in a list of 381 potential miR-375 target genes through which miR-375 may regulate glucose homeostasis. A chosen subset of these potential target genes was further investigated experimentally to establish the mechanism through which miR-375 may control glucose homeostasis.

However, selecting a handful of genes for further experimental investigation out of hundreds of potential targets is a difficult task in which the role of intuition — not to say a certain amount of luck — is not negligible. Studies whose aim is to understand the biological function and regulatory mechanism of miRNAs could greatly benefit from computational miRNA target prediction that are accurate at identifying genes likely to be efficiently regulated by the miRNA of interest out of hundred of potential target genes. Chapter 3 revisits the question of miRNA target prediction with this goal in mind. From high-throughput experimental measurement of changes in gene expression in response to miRNA perturbation experiments, we determine what additional determinants of miRNA targeting beyond seed pairing and evolutionary conservation can be taken into account in order to improve miRNA target prediction accuracy. We then explore the predictive power and the limitations of such an approach, and draw conclusions regarding the mechanism of miRNA action and the biological function of miRNAs.

So far, the datasets we have analyzed mostly studied miRNA targeting by means of perturbing miRNA expression — typically by transfecting the miRNA prepackaged in liposomes — and by subsequently measuring the regulatory consequences at the mRNA or protein levels. Such experiments are quite affordable and practically doable in cell culture. However, it has been debated whether such transfection experiments, which may result in strong, non-physiological miRNA over-expression can actually mimic the effect of miRNAs *in vivo*. In addition, such experiments, by their interventionist character, have important side-effects on cell biology which makes it difficult to distinguish genes that are directly affected by the miRNA from genes whose expression changes because of secondary effects of the miRNA transfection. Finally, the miRNA expression perturbation approach to miRNA target identification has a fundamental limitation because it focuses on

the regulatory effect of miRNAs, which takes place at the mRNA and protein level. Consequently, miRNA targets can be identified at the gene level, but determining the precise location of the miRNA binding site requires tedious additional experimental work. The PAR-CLIP method introduced in Chapter 4, which enables the identification of the binding sites of RNA binding proteins at the nucleotide resolution provides a solution to these problems. The method is applied to several RNA binding proteins, including the Ago proteins to which miRNAs provide binding specificity. As a result, genome-wide miRNA-mRNA association maps with nucleotide resolution are produced, without the need to perturb the expression of any regulator.

Chapter 5 presents MirZ, a web-based resource that makes it possible to explore miRNA-mRNA association maps and miRNA expression profiles across tissues in an integrated fashion. The rational behind MirZ is that, within a given tissue, the miRNAs that are most strongly expressed have the largest impact on mRNA targets. Therefore, deciphering the miRNA-dependent post-transcriptional regulatory layer in a given tissue or cell type has to start from the miRNA expression profile of that tissue or cell type. Conversely, it is very common that one identifies differences in miRNA expression between cells at various stages of differentiation or between normal and malignant cells, and the natural question is what mRNAs are most likely to be affected by the change in miRNA expression. The miRNA-mRNA association maps currently used in MirZ stem were obtained by computational miRNA target predictions. But ongoing software development projects in the Zavolan lab are generalizing this idea to miRNA-mRNA association maps experimentally determined by PAR-CLIP.

Finally, while PAR-CLIP provides insight into the "where" of miRNA regulation, it does not address other equally important aspects such as the time-scale on which miRNA regulation takes place or the magnitude of the regulation which can only be studied by perturbing miRNA expression. Chapter 6 shows that such kinetic aspects need to be taken into account when designing and analyzing experiments aimed at characterizing the regulatory function of miRNAs by means of miR perturbation and subsequent measurements of the induced changes in mRNA and protein levels. In addition, a detailed model of miRNA action is introduced, which makes it possible to study how different parameters influence the time-scale and magnitude of miRNA-mediated gene regulation.

2

MIR-375 MAINTAINS NORMAL PANCREATIC α- AND β-CELL MASS

The results of this chapter stem from a collaboration with the Stoffel lab at ETH Zurich, and were originally published in the Proceedings of the National Academy of Sciences USA [171]

2.1 INTRODUCTION

The maintenance of β-cell mass during development and throughout life is a highly regulated process responsible for normal glucose homeostasis. Defects in the development of pancreatic islets lead to changes in islet composition, and often result in the hyperglycemia that characterizes the diabetic state [54, 85]. The dynamic adaptation of β-cell mass in adult life is influenced by various metabolic stresses, which control the balance between proliferation and apoptosis. These processes, known to be regulated at the transcriptional level, contribute to the development and maintenance of many tissues, including the pancreatic islet [196, 108]. Recent studies have shown that miRNAs, which regulate gene expression at a post-transcriptional level, are powerful regulators of growth, differentiation and organ function [3, 12, 238]. For instance, mutant mice in which miRNAs are collectively silenced during endocrine pancreas development exhibit defects in all pancreatic lineages, including a dramatic reduction of insulin producing β-cells [144]. It is estimated that most protein coding genes are miRNA targets [69]. Combining target prediction with experimental analysis of miRNA expression and production of loss of function mutants are beginning to improve our understanding of the roles that miRNAs play in normal and disease states [238, 144, 236, 122, 210, 219]. It was previously reported that miR-375, the highest expressed miRNA in pancreatic islets of human and mice, regulates insulin secretion in isolated pancreatic β-cells [170]. In this chapter, we investigate the effect of genetic ablation of miR-375 on pancreatic islet development and function and in the etiology of type 2 diabetes.

2.2 RESULTS

2.2.1 *Development of hyperglycemia in miR-375 null mice*

To elucidate the role of miR-375 in the maintenance of glucose homeostasis and the development of the pancreatic islet in vivo, we generated miR-375 null mice (375KO) by targeted deletion and homologous recombination in ES cells. The miR-375 gene is uniquely located within an intergenic region on mouse chromosome 1, and the targeting construct was designed to eliminate the entire 64 bp miRNA precursor sequence (Figure 31A). Heterozygous mice were crossed and the mutants were confirmed by Southern blot analysis (Figure 31B). Offspring of these intercrosses revealed genotypes of expected Mendelian ratios (data not shown). An analysis of miR-375 by in situ hybridization confirmed its expression in wildtype pancreatic islets and its absence in 375KO islets (Figure 31C). Northern blotting also confirmed loss of expression in other neuroendocrine tissues in which miR-375 is expressed at low levels (Figure 31D). MiR-375 null animals are fertile

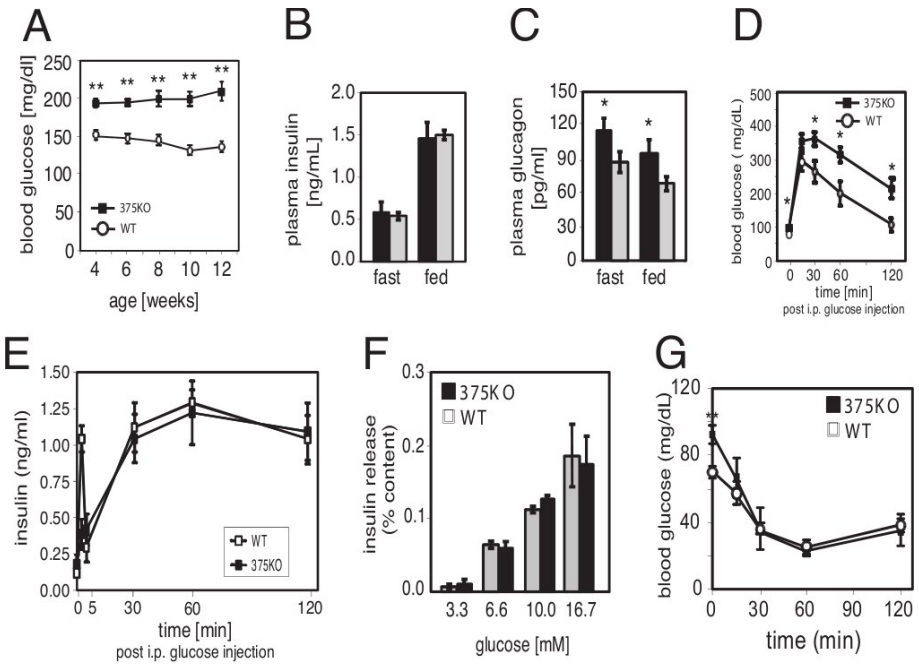


Figure 1: miR-375-null mice develop diabetes. (A) Random-fed blood glucose levels in 375KO (filled squares) and wild-type littermate control (open circles) male mice. (B and C) Plasma insulin and glucagon levels in 10-week-old 375KO mice (black bars) and wild-type (gray bars) male mice. (D) Intraperitoneal glucose tolerance test administered to 10-week-old mice. (E) Plasma insulin levels during i.p. glucose tolerance test. (F) Insulin secretion of isolated islets in response to indicated glucose concentrations. (G) Insulin tolerance test of 375KO and wild-type littermates (n=5).

and exhibit no obvious abnormalities or changes in body mass (Figure 31E).

We investigated the metabolic consequences of miR-375 ablation by measuring fed and fasted glucose and islet hormone levels. At 4 weeks, male 375KO mice exhibited random hyperglycemia (Figure 1A) and developed fasting hyperglycemia by 12 weeks (89.7 mg/dl vs. 74.7 mg/dl, $p<0.001$, 375KO vs. wildtype, respectively). Female 375KO mice developed random hyperglycemia by 8 weeks in the fed state (data not shown). Despite the hyperglycemic state, plasma insulin levels remained unchanged in 375KO mice compared to wildtype littermates (Figure 1B). In contrast, plasma glucagon concentrations were increased in both fasted and random-fed states (Figure 1C). Mutant 375KO mice exhibit elevated glucose levels compared to wildtype controls following an intraperitoneal glucose challenge (Figure 1D). Under identical conditions the first phase insulin release was diminished but plasma insulin levels were unchanged between 5 and 120 minutes after intraperitoneal glucose administration (Figure 1E). Glucose stimulation of isolated islets from 375KO and littermate control mice was similar over a range of concentrations (Figure 1F). Furthermore, no significant differences in glucose clearance were measured during an insulin tolerance test indicating the absence of peripheral insulin resistance (Figure 1G).

We have previously shown that silencing of miR-375 increases glucose-stimulated insulin secretion in pancreatic β-cell lines and isolated pri-

mary β-cells [170]. To study the effect of chronic ablation of miR-375 on insulin secretion, we therefore measured exocytosis in single, isolatedï β-cells by high-resolution capacitance measurements. Secretion was evoked by a train of depolarizations from -70 mV to 0 mV (Figure 32A). In wildtype cells, the exocytotic responses fell from an initial value of 6 fF/pF to 1.5 fF/pF at the end of the train (Figure 32B). The total increase in capacitance during the train was 34± 5 fF/pF (n=37) (Figure 32C). In β-cells lacking miR-375, the exocytotic responses fell from an initial value of 7.5 fF/pF to 3.2 fF/pF and the total response evoked by the train amounted to 55± 6 fF/pF (P<0.01 vs. wildtype; n=46)(Figure 32B,C). An identical analysis was performed on isolated α-cells, however, no differences were observed between mutant and wildtype animals (Figure 32D-F). While our earlier observations demonstrated that miR-375 is a negative regulator of β-cell exocytosis [170], these results show the hyperglycemia observed in 375KO mice is not due to a deficiency in insulin secretion.

To further analyze the underlying cause for the metabolic derangements in 375KO mice we investigated the endocrine pancreatic cell composition of mutant and control animals. Measurement of β-cell mass of 375KO pancreatic sections revealed a 38% and 31% decrease compared to wildtype controls at 3 and 10 weeks of age, respectively (Figure 2A). Quantitative morphometric analysis of 375KO pancreatic sections from 3-week old mice revealed that the change in mass was due to a comparable decrease in β-cell number (Figure 2B) and resulted in a 20% decrease in total endocrine cells per pancreatic area compared to control mice (Figure 2C). A similar decrease was observed in β-cell number at age 10 weeks in 375KO mice. In addition, these effects were accompanied by a 1.7-fold increase in α-cell number per pancreatic area compared to littermate controls (Figure 2D). The number of δ-cells was not changed in pancreata of 375KO mice compared to controls at either age (Figure 2E). No changes in total pancreatic insulin or glucagon content, or pancreatic α- and β-cell number were found at age P14 (data not shown). The results observed in 3-week old animals are the earliest detectable changes in phenotype (Figure 2A-D). The morphological analysis also revealed disrupted islet architecture with increased presence of alpha cells within the islet core and in the periphery (Figure 2F).

To investigate if elevated plasma glucagon levels could explain the hyperglycemia in 375KO mice, we evaluated glucagon secretion and downstream effects in the liver. In contrast to glucose-stimulated insulin secretion, glucagon secretion was increased in isolated pancreatic islets of 375KO mice at both low (2.8 mM) and high (25 mM) glucose concentrations compared to wildtype littermates (data not shown, Figure 2G). Furthermore, pancreatic glucagon content was increased ≃3-fold compared to wildtype littermates (375KO vs. WT: 1.25 ± 0.28 vs. 0.41 ± 0.09 ng/mg tissue, p⩽0.01, n=5). Hepatic glucose production was analyzed by measuring blood glucose levels following an intraperitoneal injection of pyruvate in random-fed mice. Significantly higher plasma glucose levels at 15 and 30 min post-injection indicated that 375KO mice have an increased ability to convert pyruvate to glucose compared with wildtype littermates (Figure 2H). In addition, 375KO mice displayed a 25% increase in the rate conversion of pyruvate-2-14C into blood glucose following of intraperitoneal injection, thereby providing further evidence that hepatic glucose production was in-

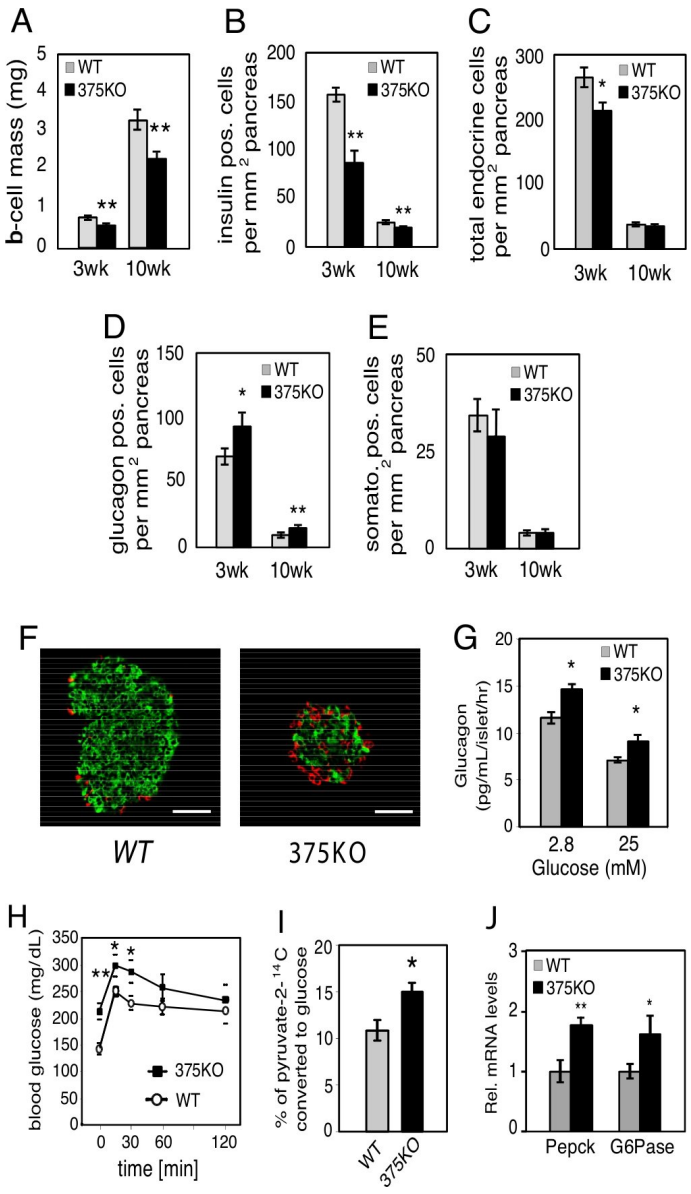


Figure 2: Decreased β-cell mass in 375KO pancreatic islets. (A) β-Cell mass in wild-type (gray bars) and 375KO (black bars) mice is quantified and reported as mean ± SE. (B-E) Quantification of endocrine cell number per total pancreatic area,β-cell number (B), total endocrine cell number per total pancreatic area (insulin, glucagon,andsomatostatin-positive cells) (C) α-cellnumber (D), and δ-cell number (E) in 375KO (black bar) and wild-type (gray bar) male mice. (F) Representative sections of pancreas from 10-week-old 375KO and wild-type male mice visualized by immunofluorescence after staining with anti-insulin (green) and anti-glucagon (red) antibodies. (Bar, 50 μm.) (G) Glucagon secretion measured from islets isolated from 10-week-old male 375KO (black bars) and wild-type (gray bars) mice cultured overnight and incubated in fresh medium containing the indicated glucose concentrations. (H) Intraperitoneal pyruvate tolerance test was performed on random-fed 6-week-old male mice by administering a dose of sodium pyruvate (in saline) at 2 g/kg body weight. (I) [2-^{14}C]Pyruvate was administered by i.p. injection into random-fed 6-week-old 375KO and wild-type (WT) mice and blood was drawn after 30 min and deproteinized, and labeled glucose in supernatant was recovered and radioactivity was measured. (J) Quantification of PEPCK and G6Pase mRNA expression by real-time PCR in liver from random-fed, 10-week-old 375KO (375KO) and wild-type (WT) mice. n = 5-12 animals per genotype unless otherwise noted. Data are presented as means ± SE. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

creased in these mice (Figure 2I). Similar results were obtained in fasted animals (data not shown), demonstrating increased de novo synthesis of glucose by the liver in both fasted and fed conditions. Moreover, under random-fed conditions, real-time PCR analysis revealed a significant up-regulation of both phosphoenolpyruvate carboxykinase (PEPCK) and glucose 6-phosphatase (G6Pase) in the livers of 375KO compared to control mice, demonstrating that the hyperglucagonemia contributes to the elevated gluconeogenesis (Figure 2J). Plasma and tissue levels of other neuroendocrine organs such as pituitary (GH, CART), adrenal (noradrenaline, adrenaline, dopamine, corticosteroids) and intestinal (GLP-1, VIP, secretin) peptides were similar in 375KO mice and littermate controls (data not shown). In addition, challenging the mice with insulin after fasting and measuring ACTH and corticosterone to test the hypothalamic-pituitary-adrenal axis revealed no abnormality between mutant and wildtype animals, indicating that loss of miR-375 expression in the pituitary and adrenal does not contribute to the phenotype of the mutant mice (data not shown). Taken together, these results show that the hyperglycemia measured in 375KO mice is primarily caused by hyperglucagonemia resulting from an increase in pancreatic α-cell mass.

2.2.2 *Expression of miR-375 is required for pancreatic β-cell compensation in obesity*

To further address the role of miR-375 in the maintenance of β-cell mass, we measured miR-375 expression in pancreatic islets isolated from ob/ob mice, a model for increased islet mass that is induced by severe insulin resistance [19]. MiR-375 expression was increased 30% in ob/ob islets compared to wildtype controls (Figure 3A). We next generated mice deficient in both miR-375 and leptin (375/ob) to determine whether the increase in β-cell mass observed in ob/ob animals is dependent upon miR-375 expression. Insulin and glucagon immunostaining from 10-week old 375/ob mice revealed an absence of islet hypertrophy compared to littermate control ob/ob mice (Figure 3B). Pancreatic β-cell mass was decreased 71% and a similar reduction was measured in total β-cell number and total endocrine cell number per pancreatic area in 375/ob animals compared to ob/ob littermates (Figure 3C-E). The relative number of pancreatic α-cells per area pancreas was unchanged in 375/ob compared to 375KO animals (Figure 3F). Consistent with 375KO mice, an increase in α-cell mass is reflected in an increase in the α- to β-cell ratio compared to both wildtype and ob/ob littermates (Figure 3G). In addition, the decrease in β-cell number in 375/ob mice was accompanied by a decrease in β-cells with Ki-67 positive nuclei (Figure 3H). No changes were observed in ob/ob mice in which only one miR-375 allele was deleted (data not shown). Failure of the islet mass to compensate for the insulin resistance induced by the obesity brought about a dramatic increase in blood glucose levels starting at age 4 weeks (Figure 3I). Consistent with decreased β-cell mass, plasma insulin levels were decreased 85% in 375/ob animals compared to ob/ob mice (Figure 3J) and plasma glucagon levels were unchanged (129.3 pg/ml ± 7.5 vs. 120.3 pg/ml ± 10.1, 375/ob vs. ob/ob, respectively, n=5-8). Furthermore, hepatic glucose production in 375/ob mice was elevated 1.8-fold compared to ob/ob mice (Figure 3K). These results, in addition to the 40% de-

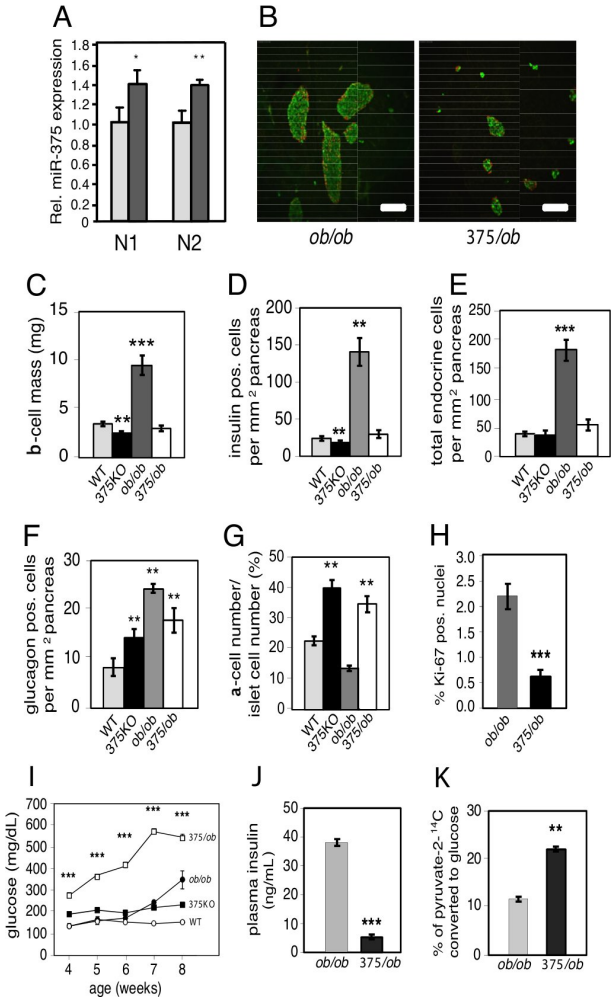


Figure 3: Impaired β-cell proliferation in miR-375/ob double-knockout mice. (A) Relative miR-375 expression in lepob/lepob (ob/ob) mice and wild-type (WT) controlsmeasuredbyreal-timePCRandnormalized toU6(N1) or miR-107 (N2) expression levels. (B) Representative 8-μmsections of pancreas from 10-week-old 375/ob (miR-375$^{-/-}$ leptin$^{-/-}$) and ob/ob mice visualized by immunofluorescence after staining with insulin (green) and glucagon (red). (Bar, 50μm.) (C) β-Cell mass in 10-week-old WT (gray bar), 375KO (black bar), ob/ob (dark gray bar), and 375/ob (open bar) mice is quantified and reported as mean ± SE. (D-F) Quantification of β-cell number (insulin-positive cells), total endo- crine cell number (insulin, glucagon, and somatostatin-positive cells) and α-cell number (glucagon-positive cells) per total pancreatic area in wild-type (gray bars), 375KO (black bars), ob/ob (dark gray bars), and 375/ob (open bars) 10-week-old mice. (G) Ratio of α-cell number to islet cell number in wild-type (gray bar), 375KO (black bar), ob/ob (dark gray bar), and 375/ob (open bar) 10-week-old mice. (H) Quantification of percentage of Ki-67 insulin-positive nuclei within insulin-positive cells of 10-week-old 375/ob (black bar) and ob/ob (gray bar) mice. n=30 for each genotype. (I) Random-fed blood glucose levels in 375/ob (open squares), ob/ob (filled circles), 375KO (filled squares), and wild-type littermate control (WT) (open circles) mice. (J) Plasma insulin levels in random-fed, 10-week-old 375/ob (black bar) and ob/ob (gray bar) mice. (K) Hepatic glucose production measured after sodium [2-^{14}C]pyruvate was administered by i.p. injection into random-fed, 10-week-old 375/ob (black bar) and ob/ob (gray bar) mice. Data are presented as means ±SE. n=4-6 animals per genotype unless otherwise noted. *, P=0.05; **, P=0.01; ***, P=0.001.

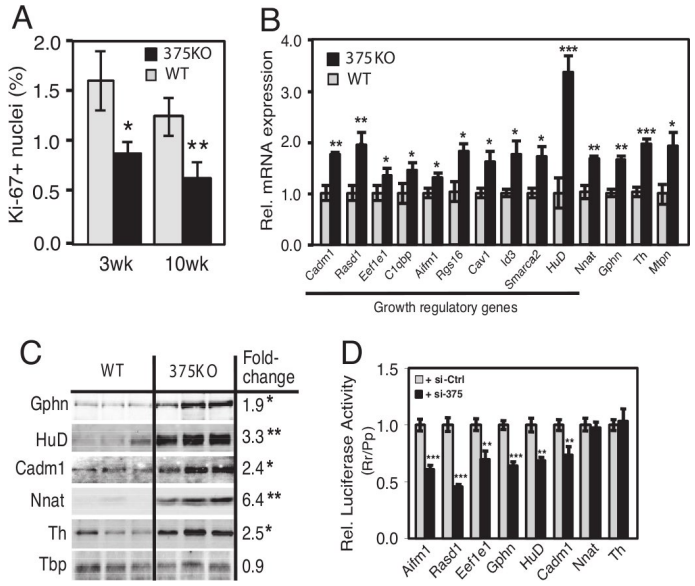


Figure 4: Regulation of gene expression and identification of growth target genes in 375KO islets. (A) Quantification of percentage of Ki-67-positive nuclei within insulin-positive cells of 375KO (black bars) and wild-type (gray bars) male mice. (B) Analysis of gene expression of putative miR-375 targets by real-time PCR in mutant and wild-type pancreatic islets. n = 5 animals per genotype. (C) Western blot analysis of protein lysates from pancreatic islets isolated from 375KO and wild-type (WT) male mice (100 islets per lane). Quantitative measurements made from densitometry are expressed as a ratio of mean values of 375KO to wild-type mice. (D) Increase in intracellular concentration of miR-375 decreases luciferase activity in HEK293 cells transfected with reporter constructs containing either full-length or partial 3'UTR sequenceof putative miR-375 target genes (n=6). Values relative to luciferase activity from cells transfected with a scrambled control are shown. Data are presented as means ± SE. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

crease in body mass and measured polydipsia and polyuria (data not shown), demonstrate severe insulin-deficient diabetes in 375/ob mice compared to ob/ob animals.

2.2.3 *MicroRNA-375 regulates genes in growth promoting pathways*

We next addressed whether the observed decrease in β-cell mass of 375KO mice could be reflective of changes in the rate of proliferation. Quantification of Ki-67-positive β-cells, an index for cell proliferation, revealed a significant decrease in 375KO islets at 3 and 10 weeks of age (Figure 4A). A similar result was obtained measuring BrdU incorporation in β-cells of 375KO mice (data not shown). To address the molecular basis for the decrease in pancreatic β-cell mass observed in the 375KO animals, we performed gene expression analysis using Affymetrix microarrays comparing tissues from mutant mice to wildtype littermates. Four tissues expressing different levels of miR-375 were selected: pancreatic islets, pituitary, adrenal, and colon. Previous studies have established that miRNAs can negatively regulate the mRNA level of their direct targets [138], and that miRNA loss-of-function can result in the up-regulation of hundreds of genes [123].

To determine the direct impact of loss of miR-375, we selected the 5% most up-regulated and 5% most down-regulated transcripts (see Supp. Methods). Each dataset thus contained 801 of the 16,301 Refseq transcripts measured by the array. We then determined the number of occurrences of the miR-375 recognition motif GAACAAA (corresponding to nucleotides 1-7 from the 5' end of the miRNA) in the 3'UTRs of these transcripts. When measuring gene expression from pancreatic islets of 375KO mice compared to wildtype littermates, we counted 138 occurrences of the miR-375 motif in the dataset of up-regulated transcripts, and 49 occurrences in the dataset of down-regulated transcripts (Figure 33A). Compared to random motifs with similar frequency across the 3'UTRs of all transcripts monitored by the array (represented in the graph by a blue box plot), the 138 occurrences represent a 1.9-fold enrichment (P=0.001), while the 49 occurences represent a 1.9-fold depletion (P=0.002). These results demonstrate that genetic ablation of miR-375 in the pancreatic islet resulted in the up-regulation of direct targets of this miRNA. To further illustrate the impact of miR-375 on islet mRNA levels, we determined the distribution of expression changes of transcripts that do include a miR-375 motif in their 3'UTR and transcripts that do not. Transcripts that carry a miR-375 motif show a significant up-regulation compared to transcripts that do not (P=2.1x10-24 in Wilcoxon rank-sum test), and the up-regulation is even stronger for transcripts containing evolutionarily-selected miR-375 motifs (P=0.005) (Figure 33E). A similar analysis of gene expression in the pituitary of 375KO mice compared to wildtype littermates revealed a significant number of up-regulated motif-containing transcripts (Figure 33C). By contrast, the genes up-regulated in the adrenal and colon data sets were not enriched for the miR-375 motif (P=0.46 and P = 0.5, respectively) (Figure 33B,D). There are two possible explanations for this discrepancy: either the magnitude of the response from direct targets of miR-375 depends upon the endogenous expression level of the miRNA, or miR-375 expression is limited to specific subpopulations of cells in the adrenal and colon. In situ hybridization using a miR-375 specific probe on pituitary tissue sections revealed miR-375 to be present in both the anterior and posterior pituitary, while its expression within the adrenal appears to be limited to the medulla and the zona glomerulosa of the cortex (Figure 34A,B). It is not known whether miR-375 is expressed in a specific cell type of the colon as probed tissue sections revealed no specific signal (data not shown).

Several genes within the set of up-regulated transcripts of miR-375 null islets have been documented to negatively regulate cellular growth and were thus evaluated for direct regulation by miR-375. Selection of transcripts that contained a miR-375 recognition motif resulted in 381 putative direct targets of miR-375. Real-time PCR analysis confirmed ten of these genes, including caveolin1 (Cav1), inhibitor of DNA binding 3 (Id3), Smarca2, Ras-dexamethasone-induced-1 (Rasd1), regulator of G-protein signaling 16 (Rgs16), eukaryotic elongation factor 1 epsilon 1 (Eef1e1), apoptosis-inducing factor, mitochondrion-associated 1 (Aifm1), cell adhesion molecule 1 (Cadm1), HuD antigen (HuD), and complement component 1, q subcomponent binding protein (C1qbp) were up-regulated in 375KO islets (Figure 4F). Increased expression of three additional genes, including cell adhesion molecule 1 (Cadm1), gephyrin (Gphn), and myotrophin (Mtpn), a previously validated target of miR-375 [170] was confirmed in 375KO islets by real-time PCR

and western blotting (Figure 4G,H). Furthermore, measurement of luciferase activity from HEK293 cells transfected with plasmid constructs containing a portion of or the entire 3′ UTR of Aifm1, Rasd1, Eef1e1, Gphn, HuD, and Cadm1 showed reduced expression of all these constructs in the presence of miR-375 (Figure 4I). These results suggest that Cav1, Id3, Smarca2, Aifm1, Rasd1, Rgs16, Eef1e1, C1qbp, HuD, and Cadm1, all of which have been shown to participate in signaling mechanisms that negatively regulate cellular growth and proliferation, are direct targets of miR-375. Published studies have shown that these genes play a role in the p53-dependent pathway [27, 72, 167], MAP kinase signaling [41], induce apoptosis [115, 37, 109], and inhibit normal developmental growth processes [179, 1] or the proliferation of tumors in mice [125, 216]. Using real-time PCR analysis, we found that the expression levels of these genes in pancreatic islets either exceed or are comparable to the levels in tissues where a functional role has previously been determined (Figure 33). We also confirmed changes in mRNA expression of several up-regulated genes that do not contain the miR-375 motif, including tyrosine hydroxylase (Th) and neuronatin (Nnat) (Figure 4H,I). While the exact role of these genes in the pancreatic β-cell is not known, it was shown that increased expression of neuronatin is associated with hyperglycemia-induced apoptosis [21, 107]. Both genes appear to be indirectly regulated by miR-375, as reporter assays with vectors that harbor their 3′-UTRs downstream of the luciferase gene did not result in decreased activity when co-expressed with miR-375 (Figure 4J). Together, these results provide evidence that many direct, as well as indirect targets of miR-375 contribute to the regulation of the β-cell composition of islets.

2.3 DISCUSSION

Our results illustrate an essential role for miR-375 in the establishment of normal pancreatic endocrine cell mass in the postnatal period and the maintenance of glucose homeostasis. The primary consequence resulting from the loss of miR-375 is chronic hyperglycemia due to a pancreatic α-cell defect, as evidenced by increased α-cell mass, increased glucagon release from isolated islets, elevated fasted and fed plasma glucagon levels, and the increase in downstream effects of glucagon such as expression of genes regulating gluconeogenesis and hepatic glucose production. Of note, 375KO mice in the fed state exhibit plasma glucagon levels that are comparable to fasted levels in wildtype mice, further emphasizing the chronic glucagon stimulus in these animals. The hyperglucagonemia in 375KO mice compared to control littermates is most likely due to the increase in α-cell number and a defect in glucose sensing since exocytosis measurements in isolated α-cells in response to direct depolarization was similar in wildtype and mutant mice. The second observation of note is that the hyperglycemic phenotype of 375KO animals is unlikely due to the decrease in β-cell mass since this reduction is usually insufficient to cause insulin deficiency and diabetes [20] and insulin secretion of isolated pancreatic islets from mutant and wildtype mice in response to various concentrations of glucose were similar. Furthermore, insulin levels in the fasted state and during a glucose challenge in 375KO and wildtype littermates also not changed, despite a reduced β-cell number in 375KO mice, suggesting that insulin secretion per β-cell is enhanced

in 375KO mice and that reduction of β-cell mass and increased secretion balance each other in mutant mice. The mechanism by which loss of miR-375 function leads to a reduced β-cell mass is most likely mediated by the cluster of negative growth regulators that are directly regulated by miR-375 and are markedly upregulated in 375KO animals. The fact that the phenotype is more profound in mice subjected to metabolic stress might indicate that miR-375 targets play a crucial role in β-cell compensation when metabolic demand is increased. The mechanism by which the α-cell number in 375KO pancreata is increased is currently unknown. Two models can be proposed: miR-375 regulates specific target genes in α-cells that are responsible for increased α-cell mass. Alternatively, the increase in α-cell number could be the result of a compensatory response to altered β-cell mass and function or to the chronic hyperglucagonemia, which in some models is associated with α-cell hyperplasia [168, 35].

Mice bearing a conditional deletion of dicer, an enzyme required for miRNA processing, during pancreas development exhibit defects in all pancreatic cell lineages, abnormal islet architecture, and a profound reduction in pancreatic β-cells [144]. Mutant 375KO mice only discreetly mimic this phenotype, suggesting that miR-375 alone is not responsible for the marked developmental defect in β-cell growth and differentiation and that other miRNAs which are expressed in endocrine pancreatic precursor cells must be responsible for the observed phenotype of the Pdx-Cre/dicer mice.

Lastly, it is interesting that miR-375 plays a significant role in the hypertrophic growth response of pancreatic islets to metabolic stress. Expression levels of miR-375 are aberrant in obese mice, indicating that they contribute to increased β-cell mass in insulin resistance. Loss of miR-375 expression in obese mice leads to a profound loss of β-cells, metabolic decompensation and premature death. Under these conditions, α-cell mass is not affected, suggesting that miR-375 has a less prominent role in α-cells, which are not under particular metabolic or cellular stress in hyperglycemic/insulin resistant conditions. Increasing evidence implicates miRNAs as an essential component mediating responses to cellular stress. For instance, tissue-enriched miRNAs in the heart, such as miR-1, miR-208 and mir-133, have been shown to regulate the hypertrophic proliferative activity in response to a variety of stresses, and miR-126 affects survival following induction of a myocardial infarction 7, 12, 33). These observations from miRNA knockout mice highlight the importance of small RNAs in cellular development, maintenance, and survival and reveal potential novel therapeutic targets for the treatment of disease.

2.4 MATERIALS AND METHODS

2.4.1 *Generation of 375KO and 375/ob mice*

The murine miR-375 gene was deleted in Sv129 embryonic stem ES cells by homologous recombination using a targeting vector in which the entire pre-miRNA was deleted and replaced by a dsRed cDNA and Neo selection cassette (Figure 31A). Targeted clones were identified by BstEII digests of genomic DNA and Southern blotting using the indicated 3'probe. Approximately 10% of clones carried the targeted allele and two clones were used to generate chimeric animals that passed

the mutant allele to offspring (Figure 31B). Double miR-375-/-, Lep-/- (375/ob) mice were generated by crossing double heterozygous mice and identified by PCR. Mice were housed in pathogen-free facilities in a 12hr light/dark cycle and were backcrossed for six generations with C57/BL6 mice before characterization of animals. The dsRed transgene was not expressed. Unless stated, male animals were analyzed at 10 weeks of age.

2.4.2 *Analysis of metabolic parameters*

Blood glucose, insulin, glucagon, free fatty acids and triglycerides in plasma were measured as described [123, 124]. Vasoactive intestinal polypeptide (VIP), cocaine and amphetamine regulated transcript (CART), and secretin were measured by radioimmunoassay (Phoenix Pharmaceuticals). The following hormones were measured by ELISA: GLP-1 (Linco), cortisol (US Biological), and growth hormone (Diagnostic Systems). Catecholamines were measured from plasma and tissues by HPLC. Individual animals were placed in metabolic cages to measure water consumption and urinary output (Columbus Instruments).

Glucose, insulin and pyruvate tolerance tests, in vivo gluconeogenesis, and HPA stimulation studies Glucose tolerance tests were performed following an overnight fast (16hr) and injected intraperitoneally with glucose (in saline) at 2g/kg body weight. Plasma glucose levels were measured from tail blood at 0, 15, 30, 60, and 120 min after infusion. Insulin tolerance tests were performed by injecting insulin i.p. (0.75 U/kg body weight), and measuring blood glucose before (time=0) and 15, 30 and 60 minutes after injection. Pyruvate tolerance tests were also performed in a random-fed state or following an overnight fast (16hr) and injected intraperitoneally with pyruvate (in saline) at 2g/kg body weight. Plasma glucose values were measured as above. In vivo gluconeogenesis studies were performed as previously described [223]. Briefly, random-fed mice were injected with sodium pyruvate-2-14C (1.5 μCi, 15 mCi/mmol) in addition to pyruvate in saline (2g/kg body weight) and 0.15 mL blood was collected via orbital sinus at 5 and 30 min. An aliquot of 0.1 mL whole blood was transferred to 0.5 ice cold water, and 0.2 mL of Ba(OH)2 and 5% $ZnSO_4$ were added in succession. After centrifugation, deproteinized blood was incubated by batch method with Amberlite Mixed Bed Exchanger MB150 resin (Sigma). Supernatants were collected and resin was washed with additional 0.2 mL water. Eluants were pooled and counted independent of separate 0.01 mL aliquots of whole blood counted to estimate the amount of labeled pyruvate absorbed into circulation. Islet secretion studies were performed on size-matched islets isolated from 10-week old animals following collagenase digestion and overnight culture and performed as described [170]. To test the hypothalamic-pituitary axis, following an overnight fast (16hr), mice received an intraperitoneal injection of insulin (0.75U/kg) and blood was taken at 0, 10, and 30 minutes post-injection. Plasma corticosterone and ACTH were measured by RIA (Peninsula Laboratories and MP Biomedical, respectively).

2.4.3 *Isolated islet secretion and capacitance measurements*

In vivo insulin release was measured in mice following an overnight fast (16hr) and injected intraperitoneally with glucose (in saline) at

2g/kg body weight. Plasma insulin was measure at 0, 2.5, 5, and 15 minutes post-injection. Islet secretion studies were performed on size-matched islets isolated from 10-week old animals following collagenase digestion and overnight culture and performed as described. Exocytosis of secretory granules was monitored in single β-cells by capacitance measurements as described previously [122, 56]. The measurements were performed in the standard whole-cell configuration of the patch-clamp technique at 32-33°C and the identity of β-cells was confirmed after the experiment by immunocytochemistry [23].

2.4.4 *Computational analysis*

The expression analysis of total RNA extracted from tissues of 10-week old animals using Trizol reagents (Invitrogen) was performed using Affymetrix mouse genome 430 2.0 arrays. Analysis of total RNA extracted from MIN6 cells infected with recombinant adenovirus expressing miR-375 as described [170] was performed using the Affymetrix mouse genome 430A array. Details on generation and analyses of data are found in Supp. Methods.

2.4.5 *Northern blotting, qPCR, immunoblotting and luciferase activity measurements*

Northern blotting, western blotting and luciferase assays were performed as previously described [170]. Antibodies for western blotting were obtained from several different sources: anti-gephyrin (Chemicon), anti-igsf4a/cadm (R&D Laboratories), anti-neuronatin (Abcam), anti-tyrosine hydroxylase (Abcam) and anti-HuD (gift of R. Darnell). For RT-PCR, total RNA was reverse transcribed using random primers according to manufacturer's protocol (Invitrogen). Primer sequences are available upon request. MiRNA qPCR results were normalized to U6 levels that were detected by using the ABI miRNA U6 assay kit (Applied Biosystems).

2.4.6 *Immunohistochemistry, islet morphometry, and in situ hybridization*

Immunohistochemistry was performed on at least five 8-µm sections (at least 160 µm apart) prepared from paraffin-embedded pancreata of 3 and 10 week old animals. Tissue sections were mounted with Vectashield with DAPI (Vector Laboratories) and analyzed using a Leica DM5500 microscope and the cross-sectional areas of pancreata and β-cells (insulin-positive cells) were determined using MetaMorph (version 7) software. Relative cross-sectional area of β-cells was determined by quantification of the cross-sectional area occupied by β-cells divided by the cross-sectional area of total tissue. β-cell mass per pancreas was determined by the product of the relative cross-sectional area of β-cells per total tissue and the pancreatic mass. Measurements were calculated by analyzing pancreata from at least 3 animals for each age and genotype. Cell quantification was based on counting nuclei of either insulin-, glucagon- or somatostatin-positive cells and data is represented as total cell number per pancreatic area. Ki-67 and BrdU-positive cells were counted from between 1500 to 2000 insulin-positive cells per animal. Antibodies for immunofluorescence were

obtained from several sources: anti-insulin and anti-glucagon (Linco), anti-somatostatin (Dako), anti-BrdU (Sigma), and anti-Ki-67 (Novocastra). BrdU incorporation and in situ hybridization was performed as described previously [236]. Specific LNA probes (Exiqon) were labeled using terminal transferase and DIG-ddUTP (Roche).

3

RELATIVE CONTRIBUTION OF SEQUENCE AND STRUCTURE FEATURES TO THE MRNA BINDING OF ARGONAUTE/EIF2C-MIRNA COMPLEXES AND THE DEGRADATION OF MIRNA TARGETS

The findings presented in this chapter were partly obtained using data from experiments performed at the Tuschl lab (Rockefeller University, New York) and were originally published in Genome Research [95]. Parts of the final discussion are to appear in the 2nd edition of the Handbook of RNA biochemistry [93].

3.1 INTRODUCTION

Since the prediction of animal miRNA targets was first tackled computationally [203, 134], many approaches, taking into account features ranging from evolutionary conservation to the position of the putative target site and the nucleotide composition of its environment, have been proposed. The constraints that functional miRNA target sites obey as well as the mechanism of miRNA action are intensely debated. The initial paradigm that emerged from the study of *Caenorhabditis elegans* miRNAs lin-4 [231] and let-7 [178] was that miRNAs induce translational repression. More recent studies challenged this paradigm and demonstrated that substantial miRNA-induced mRNA degradation occurs under both over-expression [138] as well as under physiological conditions [9]. This opened the possibility to study the determinants of miRNA targeting based on transcriptome-wide measurements of mRNA changes in response to over-expression [140, 83, 110, 191, 8, 130], knock-down [123] and knock-out [238] of miRNAs. But because the ultimate readout of the miRNA activity is the protein output of the target transcripts, the natural expectation is that measurements of protein expression changes will generate the most appropriate data sets for studying principles of miRNA-target site recognition. Such data became available very recently, when Selbach et al. [191] and Baek et al. [8] determined the changes that are induced in the proteome profiles upon miRNA over-expression and depletion by different SILAC (stable isotope labeling with amino acids in cell culture) approaches.

Extensive previous work revealed that 7-8 nucleotides at the 5′ end of the miRNA are very important for target recognition [127, 134, 47, 135, 24]. Aside from this, the sequence composition of the 3′ UTRs [182] or of the immediate environment of the putative target sites [83], the position of the site in the 3′ UTR [70, 83, 146], the base-pairing pattern in the 3′ region of the miRNA [83], the structural accessibility of the target site [182, 141, 117, 207], and the presence of multiple target sites in close proximity [59, 83] have also been reported to be predictive for the functionality of miRNA target sites. The relative importance of these features, and in particular the relative contribution of sequence versus structural determinants are at this point intensely debated.

In an attempt to identify the features that most generally characterize miRNA targets, we performed a systematic analysis of a number of large-scale publicly available data sets, each typically involving multiple miRNAs. The experiments, reported by Krützfeldt et al. [123], Linsley et al. [140], Grimson et al. [83], Karginov et al. [110], Selbach et al. [191], and Baek et al. [8], covered a variety of conditions, from miRNA over-expression to miRNA knock-down, in cell lines that expressed a normal amount of DICER1 as well as in DICER1 hypomorphs. The effects of miRNAs in these experiments were measured

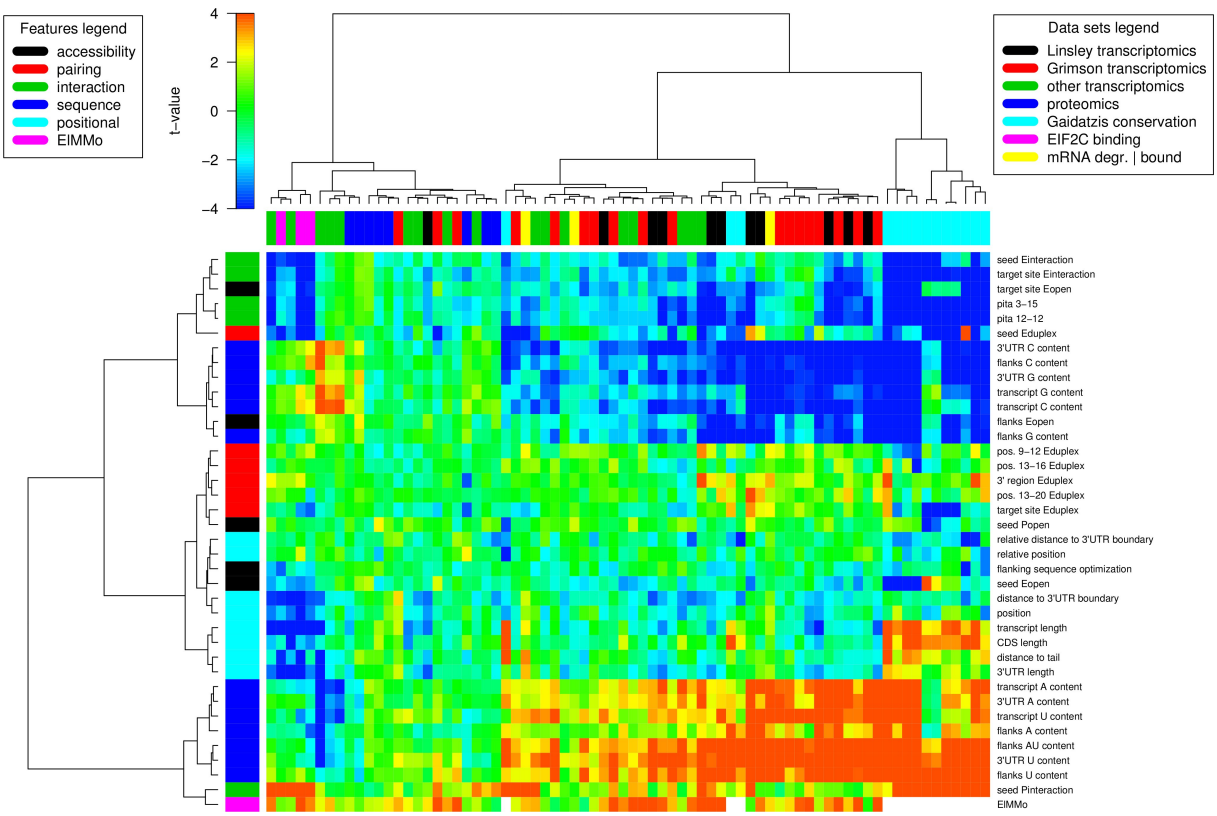


Figure 5: Predictive power of different features of putative miRNA target sites (rows) in predicting functional sites across the 74 data sets (columns). The data sets covered transcriptomics and proteomics measurements after miRNA transfection, transcriptomics measurements after miRNA knock-down, profiling of mRNAs bound to EIF2C/miRNA complexes, and target prediction based on comparative genomics. The heat-map shows the t-values comparing the distributions of feature values in functional vs non-functional miRNA target sites. The red color indicates positive predictors of miRNA functionality, while the blue color negative predictors of miRNA functionality. The dendrograms of features and data sets were produced through hierarchical clustering using Ward linkage on the euclidean space of t-values. See also Supplementary Figure 37, which also indicates the data set represented in each column.

either at the level of the transcriptome or of the proteome. In order to clarify the steps at which different features appear to come into play, we have supplemented these published data sets with our own data on transcriptome-wide changes and Argonaute/EIF2C-bound mRNAs in miRNA-transfected cells. Finally, to better understand the nature of the selection pressure on miRNA target sites, we analyzed the same set of features for sites that we previously predicted with high and low probability to be under evolutionary selection [70].

3.2 RESULTS

3.2.1 *Characterization of target sites inferred in individual studies*

The approach was to select, from each experiment, a set of functional and a set of non-functional sites, and to perform two-sample t-tests for

the difference of the mean values of various features as described in the Methods. Some of the features that we wanted to compute depend on the immediate sequence environment of the miRNA target site and we therefore only considered cases in which the mRNA-level response could be attributed with reasonable accuracy to a particular miRNA target site, for which the environment-dependent features could be unambiguously computed. Based on previous results [135, 83, 70], we thus selected the transcripts containing precisely one putative target site that matched nucleotides 1-8, 2-8, or 1-7 of the miRNA that was manipulated in the experiment. Because we found the effect of the 3'UTR sites to be more reproducible (Supplementary Figure 36) compared to that of CDS sites, we further selected those transcripts in which the putative target site was located in the 3'UTR. Finally, we only considered sites that were located at least 100 nucleotides away from the 3'UTR boundaries in order to be able to compute the environment-dependent features. The results are shown in Figure 5 (and Supplementary Figure 37), in which each feature that we computed is a row and each individual experiment is a column. The matrix cells indicate how well individual features perform in distinguishing functional from non-functional putative target sites in a particular experiment. Red and blue matrix cells denote positive and negative t-values respectively, i.e. cases in which the feature takes significantly higher (red) or lower (blue) values in functional miRNA target sites compared to non-functional target sites. For instance, the right-most column of the figure summarizes the comparison of putative miR-17 sites that have a high with those that have a low inferred probability of being under evolutionary selection [70]. The third cell from the top of that column, labeled "target site Eopen", is dark blue, meaning that the energy required to open the secondary structure of the putative target site is significantly smaller for sites with high relative to sites with low probability of being under evolutionary selection. This in turn suggests that evolutionary selection favored miR-17-complementary sites that are more accessible at the level of mRNA secondary structure. The second cell from the top of this column, labeled "target site Einteraction" is also dark blue, indicating that the energy of interaction between the miRNA and the putative target site is significantly lower (i.e. the interaction is more stable) for sites with high relative to sites with low probability of being under evolutionary selection. In contrast, the third cell from the bottom of the column, labeled "flanks U content" is dark red. This indicates that the frequency of U nucleotides is significantly higher in the regions flanking the sites with high relative to sites with low probability of being under evolutionary selection.

Applying 2D hierarchical clustering with Ward linking on the euclidean space of feature t-values reveals that the target sites inferred from most transcriptomics and from the comparative genomics data sets have similar properties. They reside in A- and U-rich sequence environments, the miRNA target region and its flanks are structurally accessible, and the binding free energy between the seed region of the miRNA and the mRNA is low. This indicates that the evolutionarily selected miRNA target sites support an mRNA degradation response to miRNAs. Strikingly, the proteomics data sets form an entirely different cluster, together with a few of the associated transcriptomics and the EIF2C (Argonaute) immunoprecipitation data sets. For this cluster the above-mentioned features are largely uninformative. This

is very surprising because the targets that were identified based on proteomics measurements are enriched in miRNA seed matches, just as the targets that were previously identified based on transcriptomics measurements [191].

One possible explanation for the less significant t-values obtained in the analysis of proteomics data sets is that the number of proteins that are sampled in the proteomics experiments is considerably lower (by a factor of 5-6) compared to the number of transcripts whose expression is measured in a microarray experiment. By scaling down the transcriptomics data sets through resampling such that we analyze similar numbers of genes from transcriptomics and proteomics experiments we found that this simple explanation does not hold (Supplementary Figure 38). On the other hand, we found that although functional sites identified in these experiments have, as expected, a higher probability of being under evolutionary selection compared to non-functional sites, the difference is less pronounced compared to that inferred from other types of experiments. This is shown in Figure 5 (feature labeled "ElMMo") for all the miRNAs covered by the proteomics experiments, and in Supplementary Figure 39 for the specific case of miRNAs that have been studied by multiple groups using a number of different technologies. This result is not due to the ElMMo algorithm having a poor ability to quantify specifically the functionality of the target sites determined through proteomics measurements, because as shown in Supplementary Figure 40, the accuracy of ElMMo in predicting proteomics data is similar to that of TargetScan context.

Although the features of functional target sites are consistent across most of the studied miRNAs, a few experimental data sets exhibit a striking reversal of the sign of the base content features, with the G and C base contents correlating positively and A and U contents negatively with site functionality. These data sets correspond to let-7 and miR-30a transfections, but not to the let-7 sites predicted based on evolutionary conservation, whose profile is consistent with that of most transcriptomics experiments. We conjecture that these observations are due to both let-7 and miR-30a inhibiting components of the RNAi pathway.

A number of studies already reported on the negative feedback that let-7 may exert on the RNAi pathway through targeting *DICER1* [67, 212] and Selbach et al. [191] already demonstrated that DICER1 protein level increases strongly (over 4-fold) upon let-7 knockdown. Similarly, we suggest that miR-30a targets the P-body component and EIF2C interactor *TNRC6A* (also known as *GW182*), which carries four matches to the miR-30a seed in its 3′ UTR, all of which are conserved all the way from human to chicken, and whose mRNA level decreases by 21% upon over-expressing miR-30a [191]. The consequence of over-expressing these miRNAs may therefore be to antagonize the effects of endogenously expressed miRNAs. Thus, the transcripts that are identified as let-7 and miR-30a targets by virtue of their down-regulation in the transfection experiments are probably transcripts that do not carry functional seed matches to miRNAs endogenously expressed in the cell, which would otherwise result in their increased expression in response to a general down-regulation of the RNAi pathway. If this were the case, we would expect the transcripts that are down-regulated in let-7 and miR-30a transfections to be depleted in functional target sites for the endogenous miRNAs. This is precisely what we found when we analyzed the transcriptomics data from Selbach et al. [191]: the

transcripts that are down-regulated in the let-7 and miR-30a transfections are significantly depleted of evolutionarily selected sites for the miRNAs that are abundantly expressed in HeLa cells (Supplementary Figure 41). One may argue that a similar behavior would be produced by the saturation/competition effect recently described by Khan et al. [118]. This effect however would apply to all the transfected miRNAs, not only to the let-7 and miR-30a, wich is not what we found. We rather observed that in the miR-1, miR-155 and miR-16 transfection experiments the mRNAs that were most down-regulated following the miRNA transfection were enriched in evolutionarily selected sites for the abundant HeLa miRNAs. The competition between the transfected and the endogenous miRNAs still occurs generally across all transfection experiments, but it is only observable at the earliest time points (Supplementary Figure 42).

To further establish that *TNRC6A* is a target of miR-30a, we cloned the *TNRC6A* 3'UTR into a luciferase vector and we transfected this into HeLa cells with or without simultaneously transfecting the miR-30a antisense inhibitor. Transfection of the *TNRC6A* reporter results in a significant reduction (48%) of the luciferase activity compared to the transfection of empty vector (Supplementary Figure 43), whereas simultaneous transfection of the miR-30a antisense inhibitor results in almost complete relief of repression. This result supports our initial conjecture that the reversal of the sign of the sequence features in the miR-30a transfection experiment is due to the negative feedback that miR-30a exerts on the miRNA pathway.

The experiments that measured the binding of EIF2C2 protein (also known as Ago2) to mRNAs resulted in the second category of data sets that exhibited the reversal of sign for the sequence features. In these data sets, the G and C contents of the transcripts and of the miRNA target site environment also correlated positively with site functionality (in this case EIF2C2 binding), while the A and U contents were negative predictors. Compared to the let-7 and miR-30a transfections, these correlations were however weaker and not significant statistically. On the other hand, structure features such as the accessibility of the miRNA binding site and the energy of interaction between the miRNA and mRNA were good predictors of the functionality of these sites, as they were for the sites inferred from transcriptomics or comparative genomics analyses. Section 3.2.2 describes our detailed investigation of the features that favor EIF2C2 binding and those that favor subsequent mRNA degradation.

Because the sequence composition of the environment of the miRNA target site affects the structural accessibility of the site, it is currently unclear which of these features is primarily undergoing evolutionary optimization. To address this question, we shuffled the sequence flanking functional miRNA target sites (keeping the miRNA target site fixed) and we asked whether the energy required to open the structure of the miRNA target site was higher in the context of the shuffled sequences compared to the real sequence. We found that for a subset of the data sets this is indeed the case (see Figure 5, row labeled "flanking sequence optimization"), providing weak but statistically significant support to the hypothesis that the sequence surrounding functional miRNA target sites is constrained to increase the accessibility of the miRNA target site beyond what can be explained from the A, C, G, U content of the flanking regions (see also Figure 6). The fact

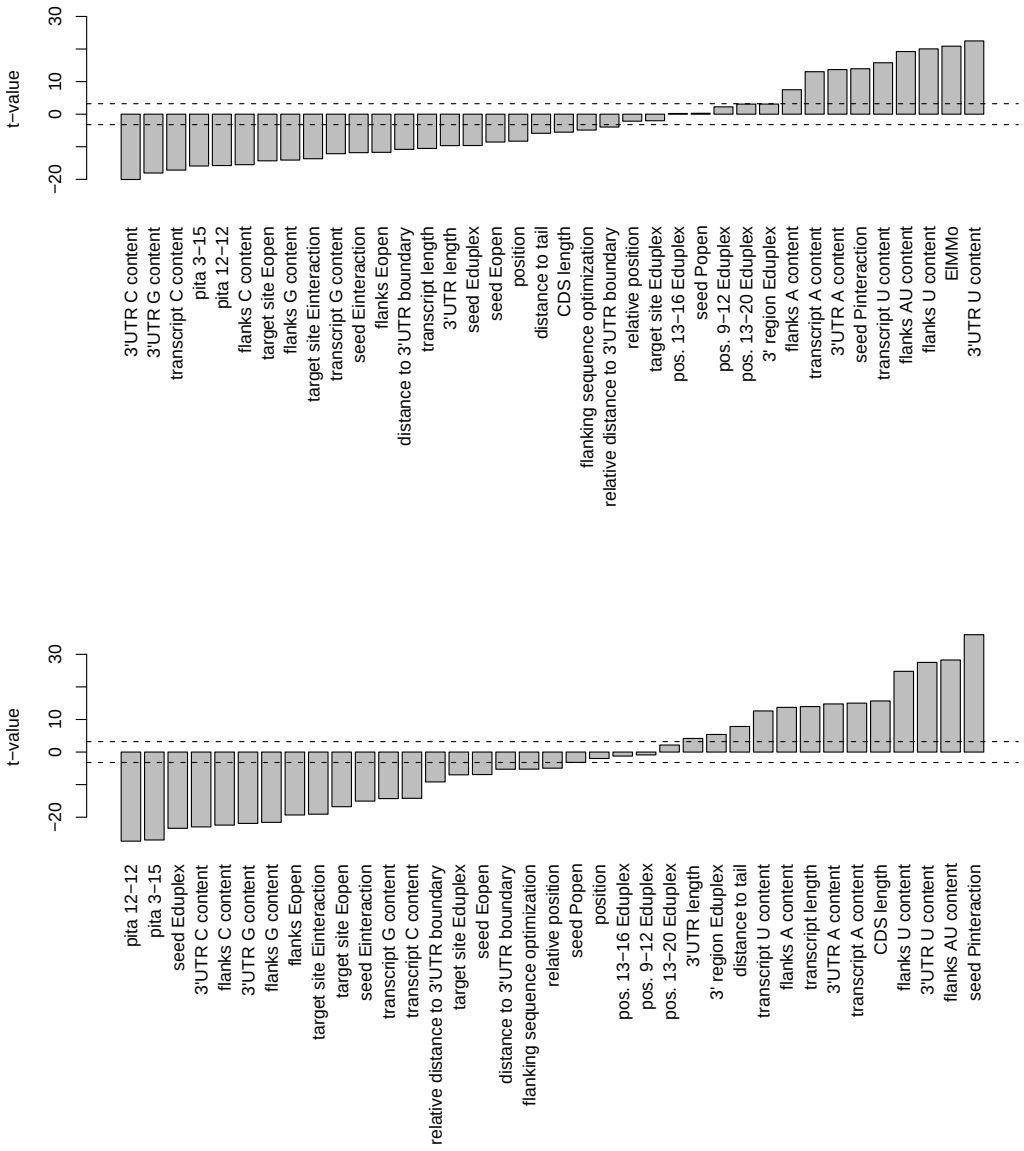


Figure 6: Upper panel: predictive power of different features across all transcriptomics experiments, excluding the let-7 and miR-30a transfections. Lower panel: predictive power of different features across all comparative genomics data sets. The y-axes show the t-values of the individual features when comparing their distribution in functional vs non-functional sites, aggregating over all data sets. The dotted horizontal lines represent the cut-off where the t-values are significant with a bilateral type I error of 5% after applying the Bonferroni multiple testing correction. Pita 12-12 and Pita 3-15 are the scores according to the algorithm described in Kertesz et al. [117], using 12-12 or 3-15 nucleotides upstream and downstream of the miRNA target site for computing target site accessibility.

that this property does not generally characterize all data sets explains in part the current controversies concerning the relative importance of sequence and structure parameters in determining the functionality of miRNA target sites [83]. We further found that with the exception of the accessibility of the miRNA flanking regions which correlates with the G+C content of these regions, the sequence features that we computed do not correlate well with the structure features (Supplementary Figure 44). This, and the results in section 3.2.2 suggest that sequence and structure features come into play in a non-redundant manner, at different steps of the RNAi effector cascade, and that it is probably necessary to take them both into account in order to understand miRNA targeting specificity.

Finally, comparative genomics-based analyses reported that miRNAs tend to target transcripts with long 3′ UTRs [204]. Strikingly, we here found that functional miRNA target sites that are identified experimentally generally reside in transcripts with relatively short 3′UTRs, and that the transcript length is an even better predictor of functionality compared to the 3′UTR length. Nonetheless, within the long 3′UTRs in which evolutionarily selected sites are found, functional sites reside closer to the 3′UTR boundaries (stop codon or polyA tail) compared to non-functional sites, as has been previously reported [70, 83, 146].

3.2.2 *Structural features direct EIF2C2 binding while sequence features are associated with mRNA degradation*

To gain insight into the origin of the sequence and structure biases discussed above, we transfected HEK293 cells stably expressing EIF2C2 with either a miRNA (miR-124 and miR-7) or a mock control, and we measured the mRNA expression in total RNA and in the RNA from the EIF2C2 immunoprecipitate (IP) with oligonucleotide microarrays. The degree of miRNA-specific EIF2C2 association and degradation of individual mRNAs were quantified by the enrichment of the respective mRNAs in the EIF2C2-immunoprecipitates and the total cellular RNA, respectively, of miRNA-transfected compared to mock-transfected cells (see Supplementary Material in chapter B). We also analyzed the results of a similar experiment performed with miR-124 by Karginov et al. [110].

Binding to EIF2C2 of transcripts whose 3′UTRs contains precisely one match to the miRNA seed was very reproducible between the two biological replicates of each transfected miRNA, with correlation coefficients of 0.85 for miR-124 and 0.70 for miR-7 (Figure 7, upper panels). Moreover, the degree of EIF2C2 binding was correlated with that of mRNA degradation (r=-0.70 for miR-124 and r=-0.62 for miR-7, shown on the the lower panels of Figure 7), with the large majority of EIF2C2-bound transcripts undergoing some degree of degradation. This correlation between EIF2C2 binding and mRNA degradation was much higher than the correlations that were reported between changes in the mRNA and in the protein levels by Selbach et al. [191] and Baek et al. [8] (see also Supplementary Figure 45). A small number of EIF2C2-bound mRNAs did not show evidence of degradation, as previously reported by Karginov et al. [110] and Hendrickson et al. [98]. To experimentally confirm that such transcripts are nonetheless regulated by miR-124 and miR-7, we generated dual luciferase reporter constructs containing the 3′UTRs of some of the EIF2C2-bound mRNAs. Cotrans-

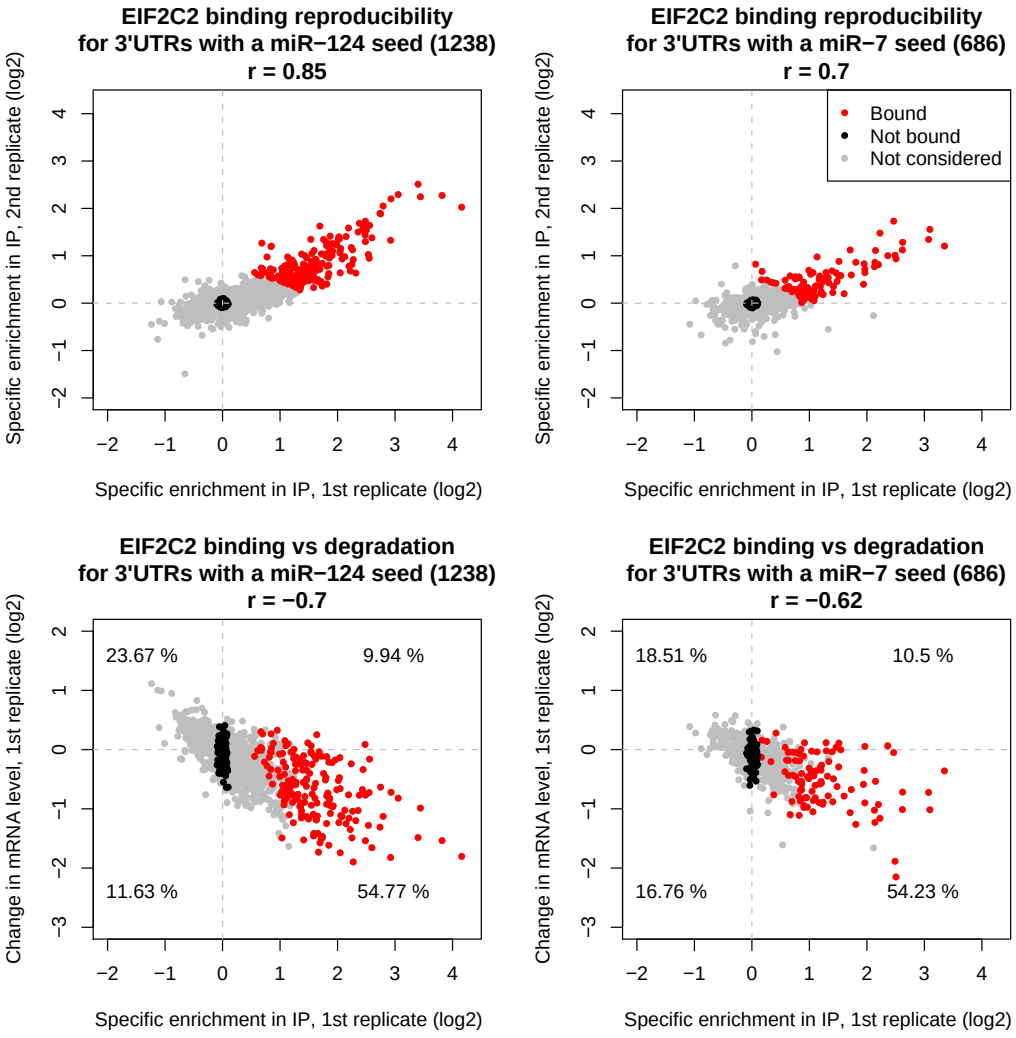


Figure 7: Upper panels: Correlation between the level of EIF2C2 binding in two replicate experiments of transcripts carrying a single seed match for miR-124 (left panel) and miR-7 (right panel) in their 3′ UTRs. The level of EIF2C2 binding was computed as described in the Methods. The number of transcripts and Pearson correlation coefficients are shown on the respective panels. Transcripts that were considered *positives* for EIF2C2 binding are marked with red, those that were considered *negatives* with black, and transcripts that were not used for feature analysis are shown in gray. Lower panels: Correlation between EIF2C2 binding and mRNA degradation in one experiment (miR-124 over-expression in the left panel, miR-7 over-expression in the right panel). The levels of EIF2C2 binding and mRNA degradation were computed as described in the Methods. The numbers in the four quadrants indicate the proportion of all transcripts with a single seed-complementary 3′ UTR site that fall in each individual quadrant.

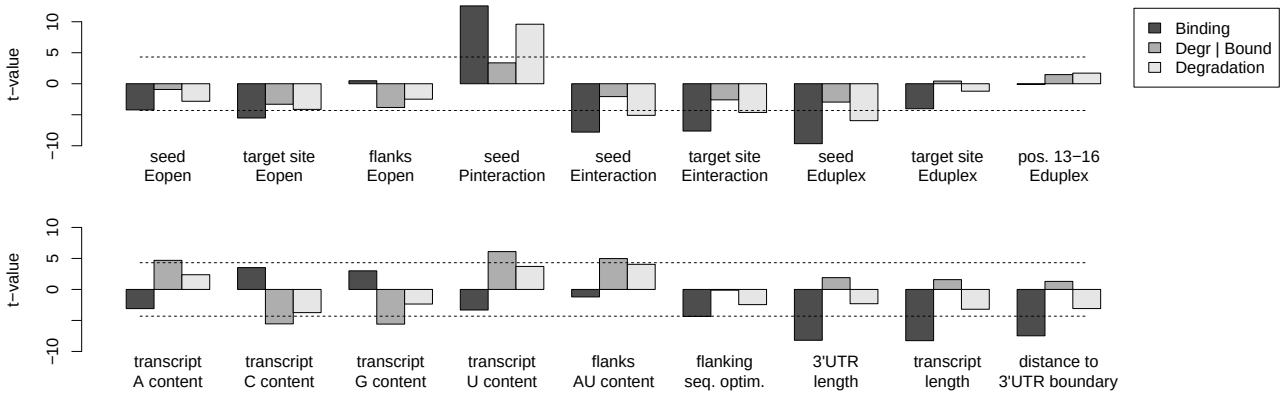


Figure 8: Contribution of secondary structure (upper panel), sequence and transcript length-related (lower panel) features to the efficiency of EIF2C2 binding and mRNA degradation. The y-axis shows the value of the t-statistic obtained in comparing bound with unbound transcripts (dark gray bars), bound and degraded with bound but not degraded transcripts (medium gray bars), and degraded with not degraded transcripts (light gray bars). The dashed lines indicate the values beyond which the difference in the mean values obtained for the positive and negative sets is considered significant with a bilateral type I error of 5% after applying the Bonferroni multiple testing correction. The individual features that we tested are indicated in the figure and further described in the text.

fections of these luciferase reporters with the respective miRNA resulted in a reduction of luciferase activity compared to control transfections indicating that irrespective of whether they undergo degradation, EIF2C2-bound transcripts are translationally repressed by miRNAs (Supplementary Figure 46).

We returned to the features that we tested on the targets inferred from all other experiments and asked at what step, mRNA binding or degradation of bound mRNAs, do these features come into play. As shown in Figure 8, the t-statistics for the energy necessary to unwind the secondary structure of the seed pairing region (labeled "seed Eopen") and of the entire target site (labeled "target site Eopen") were significantly negative, meaning that they were significantly smaller in EIF2C2-bound sites compared to unbound sites. That is, we found that 3'UTRs that are specifically bound by EIF2C2 tend to have seed- and miRNA-binding regions that are structurally more accessible, consistent with the results previously reported by Ameres et al. [4]. The energy of hybridizing the seed (labeled "seed Eduplex") makes a major contribution to EIF2C2 binding. Combining the structural accessibility of the seed-binding region with the energy of hybridizing the seed to the target site into a probability of interaction gives the most significant difference between target sites that are and those that are not bound by the EIF2C2-containing RISC complex. Note that we found the same feature to be highly predictive of miRNA sites that are under evolutionary selection (Figure 5). The 3' region of the miRNA on the other hand does not appear to play a crucial role in EIF2C2 binding to miRNA seed-complementary sites (the feature labeled "pos. 13-16 Eduplex"), consistent with the whole miRNA hybridization energy (labeled "target site Eduplex") being a weaker determinant of EIF2C2 binding than

the energy of hybridizing the seed (labeled "seed Eduplex"). None of these features however, was able to distinguish between *bound sites* that do and those that do not promote degradation.

In contrast, we found that features describing the sequence composition of the transcripts harboring miRNA target sites have a dramatic effect on the degradation of bound transcripts. While at the level of EIF2C2 binding, the nucleotides composition does not appear to play a statistically significant role, once transcripts are bound by EIF2C2, it is the U, and to a smaller extent the A content that are positive predictors of mRNA degradation (Figure 8, lower panels). The trends in nucleotide composition of the regions flanking miRNA target sites are largely a reflection of global biases. Previous studies pointed to the effect of A/U content on the efficacy of miRNA target sites [182, 106, 83], though at what step in the miRNA effector cascade this feature plays a role was so far unknown. Here we found that this feature comes into play in the degradation of EIF2C2-bound targets. We furthermore found that the U content is more predictive of functionality than the A nucleotide. Interestingly, two known examples of modulation of miRNA activity – the release of miRNA-dependent inhibition of the *SLC7A1* (*CAT-1*) mRNA by the ELAVL1 (HuR) protein under stress [17] and the inhibition of miRNA action in primordial germ cells of zebrafish by DND1 protein [114] – involve interactions of U-rich elements, and a study of mRNA decay [234] also identified a number of AU-rich elements that positively correlated with degradation rate.

Consistent with the nucleotide bias, the energy required to open the secondary structure *in the vicinity* of miRNA target sites is lower in the case of functional sites (feature labeled "flanks Eopen", Figure 5). Not all transcriptomics data sets however exhibit this property, which is probably why Grimson et al. [83] reported that secondary structure prediction was uninformative once the A/U-content of the region was taken into account. Interestingly, some of the experiments from Grimson et al. [83] (e.g. miR-133a and miR-142-3p on Supplementary Figure 37) did not show strong support for structural features, while other experiments in the same series did (e.g. miR-122 and miR-9 on Supplementary Figure 37).

3.2.3 *Implications for target prediction*

An immediate question is what features and training sets one should use in order to develop more accurate target prediction methods. To address this question, we constructed three groups of data sets:

- the transcriptomics data sets shown on Figure 5, with the exception of the let-7 and miR-30a transfections, which we left out because of the negative feedback on the RNAi pathway
- the proteomics data sets from Selbach et al. [191] and Baek et al. [8], again excluding the let-7 and miR-30a experiments
- the comparative genomics data sets from Gaidatzis et al. [70] that are shown on Figure 5

Based on a principal component analysis, we selected a set of 14 non-redundant features (listed in the legend of Figure 9) and we trained generalized linear models on the transcriptomics, proteomics and the combination of the two data sets. In the latter case, we weighted the

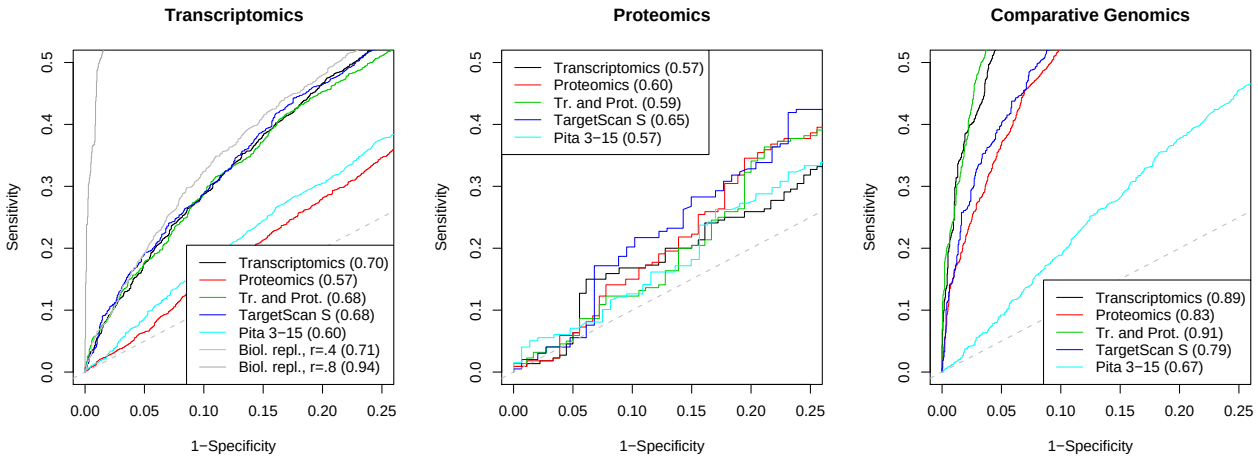


Figure 9: Receiver Operating Characteristic (ROC) curves of different miRNA target prediction algorithms on transcriptomics, proteomics and comparative genomics data sets. The numbers that appear in parentheses in the legends indicate the areas under the curves (AUC). The model fitted on transcriptomics, proteomics and combining the transcriptomics and comparative genomics data sets (Tr. and Prot.) include the following features: seed Eopen, target site Eopen, flanks G and U content, 3′ UTR length, ElMMo, seed Pinteraction, seed Eduplex, target site Eduplex, flanking sequence optimization, pos. 13–16 Eduplex, 3′ region Eduplex, distance to 3′ UTR boundary, and relative distance to 3′ UTR boundary. Pita 3-15 is the score according to the algorithm published by Kertesz et al. [117], and TargetScan is the TargetScan context score from Grimson et al. [83].

contribution of the points in the two data sets such that the combined transcriptomics measurements have equal weight in the model as the combined proteomics measurements. Because the extent of evolutionary selection measured by the ElMMo algorithm is a feature in these models, we did not train a model on the comparative genomics data only. We then assessed the predictive power of all three models on all three data sets through receiver operating characteristic (ROC) curves. We added the sensitivities and specificities of some of the current and most distinct target prediction methods for comparison. In cases where models were trained on the same data set, the ROC curve shows the cross-validation specificities and sensitivities. Finally, to get an impression of the upper bound in prediction accuracy that can be expected from a model trained on an experimental data set, we simulated duplicated experiments of varying reproducibility through sampling bivariate Gaussians with correlation coefficients of 0.4 or 0.8. This covers the range of reproducibilities found in the studies whose results we used here, such as the miR-124 transfection of Karginov et al. [110] (Supplementary Figure 47).

Unsurprisingly, each of the models that we trained performed very well on the data set on which it was trained. When it comes to predicting transcriptomics data, the model trained on these data performs as well as a replicate experiment with a relatively low (0.4) correlation coefficient would perform (Figure 9, left panel). In other words, given the noise in some experimental data sets, it is not possible to train a better model from these data, although the situation may change as more reproducible data sets become available. This is illustrated by the comparative genomics ROC curves (Figure 9, right panel), where it is possible to achieve areas of the curve (AUC) of 0.91, while no model is able to achieve an AUC greater than 0.7 on the transcriptomics or proteomics data sets. On the proteomics data set even the model trained on proteomics only achieves an AUC of 0.6, suggesting that either entirely novel features have to be taken into account in order to explain the protein-level changes that are induced by miRNAs, or that these data sets are too preliminary for studying the determinants of miRNA targeting specificity. Overall, the model trained on transcriptomics data generalizes very well to comparative genomics data, though additionally training on the proteomics data still improves slightly the prediction accuracy. Of the previously published models, TargetScan context has good performance on all data sets, which is perhaps due to the fact that it uses features that were inferred from both comparative genomics as well as miRNA transfection and transcriptomics analysis. Interestingly, its performance on the proteomics data set is better even than the performance of the linear model that we trained on the proteomics data itself. On the other hand, the performance of the Pita algorithm [117] suggests that attempting to predict miRNA targets purely from secondary structure considerations is currently not an optimal strategy.

3.3 DISCUSSION

3.3.1 *A model that combines both sequence as well as structural aspects performs best in miRNA target prediction*

The studies of the determinants of miRNA targeting specificity that have been published so far can be divided into two main classes: those that emphasize sequence features [182, 83], and those that emphasize mostly structural aspects [182, 4, 141, 117, 91]. Because different studies used different systems, looked at different readouts and had different degrees of precision in the experimental measurements, it has been difficult to reconcile their conclusions concerning the relative importance of these feature in the prediction of miRNA target sites. Here we addressed this problem by applying a uniform battery of tests in order to determine the relative power of individual features in distinguishing functional from non-functional target sites. The general conclusion is that a model that combines both sequence as well as structural aspects performs best in miRNA target prediction. The features have nonetheless to be carefully chosen, because the physico-chemistry of miRNA-target interactions is not well characterized at the moment. Thus, although the energy of interaction between a miRNA and its target is generally not a very good predictor, especially when one does not specifically enforce the hybridization of the miRNA seed, structural descriptors improve the predictive power of models that are only based on sequence features. Of the sequence features, we found that the U and A/U content of the 3'UTRs are the strongest positive and the C and G content of the 3'UTRs are the strongest negative predictors of miRNA target site functionality (Figure 6). The question arises of why nucleotide biases computed over regions of the length scale of 3'UTR lengths are predictive of the functionality of individual sites. One possible answer is that the entire 3' UTR contributes to the accessibility of individual miRNA binding regions. Consistent with this hypothesis we found that miRNA target site accessibility is one of the strongest structural predictors of target site functionality. On the other hand, we found that target site accessibility is only important for EIF2C2 binding, for which a high A/U content is not predictive. Another possible answer is that various selection pressures act to optimize the nucleotide composition over relatively long regions of the 3'UTR. This is consistent with the idea that transcripts of certain functional categories such as transcription factors, are heavily regulated [181, 204], and as a result their 3'UTR are docking platforms for a multitude of regulatory factors all of which prefer structurally accessible regions. An interesting implication of the length scale of nucleotide compositional biases is that functional target sites will more likely emerge in 3'UTRs that already have such sites, accompanied by a specific nucleotide bias that extends over long regions. A final possibility is that efficiency of mRNA degradation by exonucleases depends on how extensive the secondary structure of the transcript is. In this scenario, the A/U content of the transcript and its 3' UTR is not an indicator of the functionality of a miRNA site *per se*, but sites that are located in A/U-rich transcripts are associated with more efficient target mRNA degradation.

3.3.2 *miRNA target sites have been selected in evolution on their ability to trigger mRNA degradation*

The original paradigm regarding the mechanism of action of miRNAs was that miRNAs cause translational repression of bound mRNAs [231, 178]. Further studies have then shown that miRNAs also trigger the degradation of the targeted mRNAs [138, 9, 123], leading to the view that miRNAs primarily cause translation repression, with mRNA degradation occurring as a by-product [61]. Our results here show that the target sites that are under evolutionary selection share most features with the target sites that induce mRNA degradation responses. Thus, we suggest that the translational inhibition only paradigm is the exception rather than the rule at least for mammalian miRNAs. This conjecture is also supported by the results of EIF2C2-IP and miRNA over-expression/proteomics experiments. The degree of EIF2C2 binding correlates very well with the extent of mRNA degradation (Figure 7) and there are relatively few targets that appear to be bound by EIF2C2 but not undergo mRNA degradation. Additionally, the proteomics data sets of Selbach et al. [191] also indicate that there are relatively few targets that appear to be translationally inhibited yet the corresponding mRNA levels are unchanged (Supplementary Figure 45). One important exception may be those mRNAs whose translation needs to be inhibited only transiently. Bhattacharyya et al. [17] described for instance the example of the cationic transporter (CAT-1) message, whose inhibition by miR-122 in the liver is reversible under stress. Similar situations arise in neurons, in which the translation of some messages needs to be specifically triggered in response to signals at the neuronal synapse, but not otherwise. For such cases, the measurement of protein levels may be essential in target identification, and it will be extremely interesting to analyze in more depth the targets obtained from proteomics and from transcriptomics measurements performed after transfection of the neuron-specific miRNA, miR-124. Nonetheless, our results indicate that the more common transcriptomics measurements are still very useful for the identification of miRNA targets.

Finally, we found that a model that was trained on transcriptomics data performs better in predicting target sites that are under evolutionary selection than those that are inferred from transcriptomics experiments and that miRNAs appear to feed back on various steps of the RNAi pathway. These findings suggest that a more accurate identification of miRNA target sites may require a deeper quantitative understanding of the miRNA-induced response rather than additional determinants of miRNA targeting specificity. We will elaborate on these aspects in Chapter 6.

What we have not addressed up to this point are the practical aspects of using miRNA target predictions in the framework of a specific biological question. We have also not elaborated on some of the steps involved in designing miRNA target prediction algorithms that are not straightforward and are dependent on aspects of gene regulation which are presently only partially understood.

In this section, we will address these questions. We will start by examining issues that arise when using miRNA target predictions for answering specific biological questions in an experimental setting. Then, we will discuss how some of the uncertainties regarding gene regula-

tion are reflected in computational miRNA target predictions and to what extent one can deal with these uncertainties.

3.3.3 *Using miRNA target predictions in an experimental setting*

A typical setting in which miRNA target predictions are useful is when the miRNAs that are involved in a specific process have been identified by miRNA expression profiling or genetic screens, and the question becomes what targets respond to these miRNAs. Target predictions are then necessary to guide target discovery. This approach has been used in numerous studies aiming to understand, the role of miRNAs in development [231, 131, 178], insulin secretion [170], cholesterol biosynthesis [123], or pathologies [97, 185], to mention only a few cases.

HOW ACCURATE ARE MIRNA TARGET PREDICTIONS? If miRNA target prediction is essential for the identification of miRNA targets, the immediate question is what target prediction program should an experimental biologist use. The literature does not provide a clear answer to this question for a number of reasons. First is that it is still unclear what experimental data is suitable to assess the quality of miRNA target predictions. Are measurements of mRNA stability sufficient, or does one really need measurements of the protein levels in order to identify miRNA targets? Second, given that miRNAs appear to act at multiple levels, it is unclear how we should treat the predictions. Should one require that the response to miRNA perturbations is predicted quantitatively or would it be sufficient to predict whether the target responds or not? Another complication in comparing target prediction methods is that there is no standard set of potential targets that are always considered. That is, some authors used Refseq transcripts, others use transcripts from specialized databases (such as WormBase [92] or FlyBase [213]), even the content of a given database (such as Refseq) changes in time, and thus one method may have a prediction that another is missing simply because the transcript was not even considered. Nonetheless, we attempted to perform such a comparison on a subset of the available, commonly used, miRNA target prediction programs and a set of genes and transcripts that were used in all of these programs. Our results [95] indicated that the there are a number of miRNA target prediction programs that perform comparably well and explain around 20% of the variance in the changes in gene expression induced in a miRNA perturbation experiment. This means that 1) there is no clear "best miRNA target prediction method") and 2) the predictive power of the best miRNA target prediction methods available today is comparable to that of duplicated experiments of low reproducibility [95]. One category of targets that is generally not predicted by the methods that are currently available is that of the so-called "non-canonical" miRNA binding sites, that is, sites that cannot extensively pair with the miRNA seed. While such sites were in fact among the first to be identified [178], attempts to predict them presently come at the cost of a dramatic loss of specificity [135, 70] and, on average, their impact on the mRNA stability appears to be limited *in vivo* [90].

One thing that we need to keep in mind though is that high-throughput experiments of miRNA perturbations were frequently done on (cancer) cell lines. It is likely that in such systems the miRNA effects are more

easily interpretable compared to in vivo situations, and moreover, that miRNA transfections usually lead to large changes in miRNA expression that may not be typical of the *in vivo* situations. Additional pitfalls arising from the use of high-throughput datasets for assessing the accuracy of miRNA target predictions are discussed in the second part of section 3.3.4 below. To circumvent the current limitations of high-throughput methods, one can focus instead on high-quality, experimentally validated "positives" and "negatives" [166], though the set of such targets is significantly smaller and perhaps not even representative for the entire set of miRNA targets.

"WHICH MIRNA TARGET PREDICTION METHOD SHOULD I USE?" Because no method can currently predict miRNA targets very accurately, some authors attempted to obtain high-confidence predictions by intersecting the results of several prediction methods. While this idea may sound reasonable in theory, it may not necessarily result in more accurate predictions in practice, as can be illustrated by a simple example.

Let us imagine that we are given the list of genes that are predicted to be targeted by a certain miRNA by two different methods. Let us further assume that method A is a very accurate prediction method, while method B simply consists of tossing a coin for each gene in the human genome, calling the gene a predicted target when the coin toss returns heads. Intersecting the two lists of "predictions" we obtain a "random" subset of the targets predicted by method A. The fraction of real targets within this subset (the positive predictive value) will remain the same as for method A. The fraction of real targets that will indeed be predicted as targets (the sensitivity) in the intersection will be half of that in the list of predictions from method A alone. Thus, the accuracy of this method that combines prediction lists is lower than the accuracy a single method (method A). In fact, the situation may be even worse, when the current best method is not among those whose prediction lists will be intersected. Thus, in the context of an experiment-driven project, when the aim is to find what genes regulated by a given miRNA explain a certain phenotype, one should rather start by considering the assumptions made by the different prediction methods that are available. For instance, if the miRNA of interest is conserved in evolution we may expect that its targets are also conserved and we could consider miRNA target prediction methods that rely on sequence conservation. If the miRNA is itself poorly conserved, one should rather consider methods that aim to predict miRNA-complementary sites that induce mRNA destabilization, even though it is presently unclear whether these methods have higher or lower accuracy relative to comparative genomics-based methods.

HOW MANY TARGETS DOES MIRNA X HAVE? One of the main aims of miRNA target prediction is to generate a list of genes sorted in descending order of confidence of the gene being a miRNA target. Depending on the method and on the miRNA, one typically finds that the number of predicted targets ranges from a dozen to thousands of genes. Therefore, a natural question arising from such lists is: how many targets does a miRNA have?

This is again a difficult question to answer in principle, not in the least because, as we discussed above, it is not entirely clear how to

define a miRNA target. From the point of view of computational predictions, some miRNA target prediction methods propose criteria that are typically based on statistical considerations (signal to noise ratio, a posteriori probability, etc.) to decide where to cut off the list of predicted targets to be considered for experimental validation. With this approach it has been inferred that a miRNA targets on average hundreds of genes, the number varying between a couple and thousands of genes [134, 138, 8, 191, 69] for individual miRNAs.

WHY DOES A PARTICULAR HIGH-CONFIDENCE PREDICTED TARGET NOT CHANGE IN RESPONSE TO MIRNA OVER-EXPRESSION? It is not uncommon to find out that a substantial fraction of high-confidence predicted targets do not respond in a particular validation experiment. Of course, a trivial possibility is that the prediction is erroneous. In section 3.3.4 we will discuss several scenarios in which the target cannot be validated experimentally even though it is indeed a target.

TRANSCRIPT X IS A TARGET OF MIRNA Y ACCORDING TO METHOD Z, YET IT DOES NOT HAVE A "MIRNA Y SEED MATCH" IN THE 3'UTR. An frequently overlooked cause for discrepancies between target prediction methods or between the predictions and validation experiments is that the sequences associated with specific transcript identifiers change between database releases. Thus, predicted miRNA targets that have been downloaded from the web sites associated with specific methods may differ in sequence from the transcripts that one can download at a later date from databases such as NCBI, Ensembl or DDBJ.

3.3.4 *The complexity of gene regulation and its impact on designing accurate miRNA target prediction methods*

In order to establish that there is a direct interaction between a miRNA and an mRNA target, one possibility consists in perturbing the miRNA binding site by deletion or mutation. This approach was largely used to understand the fundamentals of miRNA target recognition [24], and is well suited for confirming a small set of putative targets experimentally. However, this approach is not practical for the identification and validation of all target sites of a miRNA. For this purpose, the alternative approach that consists in perturbing the miRNA by over-expression, knock-down or mutation was widely used to find what genes are targeted by miRNAs [123, 8, 191]. The down-side of such experiments is that the perturbation of the miRNA will percolate through the regulatory networks of the cell, which respond at different time scales, which in the end will complicate the interpretation of the perturbation experiment. This is one reason why for instance, not all mRNAs that are down-regulated following miRNA over-expression harbor a miRNA seed match [138]. "Secondary" effects superimpose with the "direct" miRNA-induced gene silencing effects, complicating the selection of "positive" and "negative" binding sites that one would use to train the predictive model. We will review here briefly the scenarios in which an incorrect selection of "positive" or "negative" binding sites is made, particularly those due to such secondary effects.

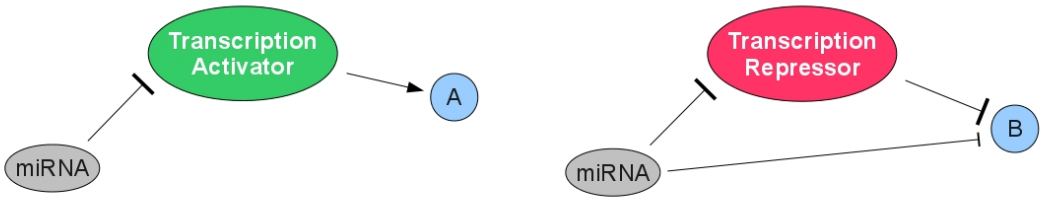


Figure 10: Hypothetical networks illustrating the co-regulation of a gene (A, B) by a miRNA and a transcription factor, with the miRNA regulating the transcription factor. Gene A harbors a non-functional miRNA binding site while Gene B carries a functional miRNA binding site.

The most obvious causes for an incorrect identification of positives and negatives are the technical measurement error (noise) and the intrinsic biological variability. Such errors can be reduced by improving the technology and performing a larger number of replicated measurements. Another obvious factor is the concentration of the reactants (mRNA, miRNAs). The target mRNA needs to be expressed at a level that allows its detection and an accurate estimation of the change in response to the miRNA. If these conditions do not hold for the cell line used in the miRNA perturbation experiment, the mRNA will not be identified as a target. The concentration of the miRNA matters as well: knocking out a miRNA which is only present in trace amount or over-expressing a miRNA whose expression already very high in the cell line in which the experiment is performed is unlikely to produce a measurable change in the expression of the targets.

The "secondary effects" come from the interaction of the perturbed miRNA with the gene regulation network of the host cell. Because regulatory networks have not been sufficiently characterized quantitatively, one cannot simply predict the secondary effects of a miRNA over-expression. Nonetheless, given that miRNAs and transcription factors co-regulate targets, with the miRNA often regulating the transcription factor or the transcription factor regulating the miRNA [192], such effects are expected to be important. Consider the example of a hypothetical experiment in which one over-expresses a miRNA that silences a transcription factor which actives the transcription of mRNA A. We assume that A harbors a non-functional miRNA binding site (Figure 10). By inferring "positive" miRNA binding sites from mRNAs whose levels go down following the miRNA over-expression, one would treat mRNA A as a functional target of the over-expressed miRNA (a false "positive"). On the other hand, we can consider the hypothetical case of a miRNA which silences a transcription factor that represses a mRNA B. Let us further assume that mRNA B harbors a functional binding site for the over-expressed miRNA (Figure 10). Over-expressing the miRNA will lead to a down-regulation of the transcriptional repressor which could then result in an up-regulation of mRNA B despite the direct regulation by the miRNA. In this case, from the measured changes in expression, we would treat mRNA B as non-functional miRNA target (a false "negative"). Other known examples of such secondary effects include the targeting of components of the miRNA pathway itself such as Dicer [67, 191] or TNRC6 [95], or of enzymes involved in chromatin structure [50, 198] by the perturbed

miRNA. In addition, the over-expression of a miRNA has been shown to impact the post-transcriptional regulatory network of a cell through competition with the endogenously expressed miRNAs [118].

Other post-transcriptional regulatory mechanisms (that may be triggered in the miRNA perturbation experiment) have been reported to interfere with miRNA regulation. For instance, under stress conditions RNA binding proteins can relieve miRNA-mediated translational repression [17], and miRNAs may even switch from acting as repressors to acting as activators [220]. Along the same lines, the RNA-binding protein Dnd1 can relief miRNA-dependent inhibition by blocking the access of the miRNA to its site [113]. Thus, there are a variety of factors that can change the functionality of miRNA target sites in a context-dependent manner.

Strongly expressing a transcript with multiple sites complementary to a given miRNA was shown to derepress the targets of that miRNA [53, 75] which suggests that miRNA targeting is transcriptome-dependent; in the context of a certain transcriptome, a miRNA binding may be functional, but in the presence of another strongly expressed mRNA that recruits most copies of the miRNA, the same target may be derepressed and therefore appear non-functional [190, 7]. As the transcriptome largely depends on the tissue and the experimental conditions, taking into account the cell type-specific transcriptome may lead to more accurate identification of positive and negative sites and more accurate miRNA target predictions.

Finally, although it is generally accepted that miRNAs silence their target genes both through repressing the translation of the transcript and by promoting deadenylation leading to the degradation of the target mRNA [66, 64], it can not be excluded that a set of target mRNAs are only repressed translationally without undergoing degradation [95]. For these sites, the transcriptomics approach is bound to generate false "negatives".

These examples illustrate that obtaining sets of positive and negative miRNA binding sites from miRNA perturbation and omics experiments is not trivial and is necessarily associated with a certain amount of "noise". The advantage of the large amount of data produced by omics experiments is that this "noise" is expected to cancel out if enough sites from different experiments with different miRNAs are used to train miRNA target prediction algorithms. The fact that miRNA target sites obtained from heterogeneous sources share similar properties argues for this scenario [95].

The discussion above also suggests that there are miRNA targets that have been predicted computationally but have not been validated experimentally because in the experimental context used to validate the target, the real effect of the miRNA on the putative target is masked by one or factors mentioned above. Reconciliating miRNA target predictions with validation experiments will require to take the precise context of the validation experiments into account.

3.4 METHODS

microRNA transfection

FLAG/HA-EIF2C2 cells were transfected with miR-7/miR-7* duplex (5′-UGGAAGACUAGUGAUUUUGUUGU/5′-CAACAAAUCACAGUCUGCCAUA)

and miR-124/miR-124* duplex (5′-UAAGGCACGCGGUGAAUGCCA/5′-CGUGUUCACAGCGGACCUUGA) or mock and Lipofectamine RNAiMAX as described by the manufacturer (see also Supplementary Figure 48). Briefly, 15 cm tissue culture plate was transfected with 900 pmol miRNA duplex and 22 µl Lipofectamine RNAiMAX.

RNA isolation from cell lysate and FLAG-protein immunoprecipitates from FLAG/HA-EIF2C2 expressing cells were lysed in 3 cell pellet volumes of 50 mM HEPES-KOH, pH 7.4, 150 mM KCl, 2 mM EDTA, 0.5 mM DTT, 1 mM NaF, and 0.5% NP-40. RNA from the lysate was isolated by adding 3 volumes of RNA extraction solution (4 M guanidinium isothiocyanate, 25 mM Na-citrate, 0.5% N-Lauroylsarcosinate, 50 mM beta-mercaptoethanol and 50% acidic phenol) and 0.2 volumes of chloroform. RNA was ethanol-precipitated from the aqueous phase. FLAG/HA-tagged EIF2C2 was immunoprecipitated with anti-FLAG M2 agarose beads (Sigma). Beads were washed three times with 50 mM HEPES-KOH, pH 7.4, 300 mM KCl, 2 mM EDTA, 0.5 mM DTT, 1 mM NaF, and 0.05% NP-40. RNA isolation from immunoprecipitated RNPs was performed as described previously [153]. RNA for microarray analysis was further purified using RNeasy mini spin columns (QIAGEN). Quality of the RNA was assessed with an the Agilent Bioanalyser.

Dual Luciferase assay of EIF2C2-bound mRNAs

HEK293 cells were co-transfected in 96-well format (40.000 cells/well) with 100 ng of the respective psiCHECK vector and 10 pmoles of miRNA duplex or 10 pmoles of GFP siRNA duplex (5′-GGCAAGCTGACCCTGAAGTTT/5′-ACTTCAGGGTCAGCTTGCCTT) as control with Lipofectamine 2000 (Invitrogen). Cells were lysed in 1xPassive Lysis Buffer (Promega) 15 h after transfection and analyzed using the Dual-Luciferase Reporter System (Promega) as described by the manufacturer on a BIO-TEK Clarity 96-well plate reader with double injectors.

Dual Luciferase assay of TNRC6A with miR-30a

HeLa cells were transfected in 24-well plates with 5 ng of respective psiCHECK vector using Lipofectamine 2000 (Invitrogen) or co-transfected with 20 nM miR-30a antagomiR (Ambion). Cells were lysed 24 h after transfection and luciferase activities were measured using the Dual-Luciferase Reporter System (Promega) as recommended in the manufacturer instructions.

Microarray experiments

Two µg of purified total RNA from HEK293 cell lysate or from immunoprecipitated RNPs were used in the One-Cycle Eukaryotic Target Labeling Assay (Affymetrix) according to the manufacturer's protocol. Biotinylated cRNA targets were then cleaned up, fragmented, and hybridized to Human Genome U133 Plus 2.0 Array (Affymetrix).

Computational analysis of one-channel Affymetrix microarrays from Selbach et al. [191] and Krützfeldt et al. [123]

The CEL files of Selbach et al. [191] were downloaded from http://psilac.mdc-berlin.de/download/ and the antagomiR-122 data of Krützfeldt et al. [123] was retrieved from the GEO database of NCBI (accession: GSE3425).

We imported the CEL files into the R software (www.R-project.org) using the BioConductor affy package [74]. The probe intensities were corrected for optical noise, adjusted for non-specific binding and quantile normalized with the gcRMA algorithm [233].

Per gene log2 fold change were obtained through the following procedure. We first fitted a lowess model of the probe log2 fold change using the probe AU content. We used this model to correct for the technical bias of AU content on probe-level log2 fold change reported by Elkon and Agami [57]. Subsequently, probe set-level log2 fold changes were defined as the median probe-level log2 fold change. Probe sets with more than 2 probes mapping ambiguously (more than 1 match) to the genome were discarded, as were probe sets that mapped to multiple genes. We then collected all remaining probe sets matching a given gene, and averaged their log2 fold changes to obtain an expression change per gene. For sequence analyses, we selected for each gene the RefSeq transcript with median 3′ UTR length corresponding to that gene.

Finally, we considered all genes for which at least one probeset was called present in the transfection experiments as expressed, and went on analyzing only these genes while ignoring all other genes.

Computational analysis of two-channel Agilent microarrays from Karginov et al. [110] and Baek et al. [8]

The Baek data set was downloaded from the GEO database of NCBI (accession GSE11968). For the Karginov data set we started with the text file output of the Agilent scanner, which was kindly provided to us by Ted Karginov.

We extracted the rProcessedSignal, gProcessedSignal, LogRatio, rIsWellAboveBG, gIsWellAboveBG fields for each probe, keeping only probes for which both gIsWellAboveBG and rIsWellAboveBG flags were true in all experiments. We then quantile-normalized the green and red channel intensities which we obtained from the rProcessedSignal and gProcessedSignal fields of all experiments together. We computed probe-level log2 fold changes from the quantile-normalized rProcessedSignals and gProcessedSignals.

After discarding probes mapping to multiple genes, we collected all probes matching a given gene, and we estimated the log2 fold change per gene as the average log2 fold change of the probe sets associated with it. Finally, for each gene we selected for further sequence analysis the RefSeq transcript with median 3′ UTR length corresponding to that gene.

Computational analysis of two-channel Agilent microarrays from Linsley et al. [140] and Grimson et al. [83]

We downloaded the processed differential expression data from GEO (accessions: GSE6838 and GSE8501) together with the probe to tran-

script mapping provided by the authors as a SOFT formatted file. For subsequent analysis, we kept only probes associated to RefSeq transcripts according to the annotation. We used all the experiments in the Grimson et al. [83] series. From the microarray data provided by Linsley et al. [140], we kept only those experiments that had quasi-replicates (transfections in both HCT116 and DLD-1 cells). These involved let-7c, miR-103, miR-106b, miR-141, miR-15a, miR-16, miR-17, miR-192, miR-200a, miR-20a and miR-215 transfections and microarray measurements at 24h (GEO accessions: GSM156546, GSM156550, GSM156545, GSM156549, GSM156543, GSM156576, GSM156532, GSM156541, GSM156534, GSM156542, GSM156580, GSM156544, GSM156547, GSM156551, GSM156548, GSM156552, GSM156553, GSM156555, GSM156554, GSM156556, GSM156557, GSM15655).

Computational analysis of SILAC assay from Baek et al. [8]

We downloaded the data provided by the authors in the supplementary material of the paper and used it without any specific post-processing.

Computational analysis of pSILAC assay from Selbach et al. [191]

We downloaded the "all peptide evidence" flat file from http://psilac.mdc-berlin.de/download/.

We mapped all peptides in the pSILAC data set against the RefSeq Protein database from Aug, 14th 2008 using wu-blastp 2.0 and a seed word length of 5, discarding alignments with gaps or with more than one mismatch. We further discarded peptides that mapped to more than one protein.

Per-protein log2 fold changes were computed for all proteins credited with 3 to 15 peptides log2 fold changes across replicates and gel slices.

EIF2C2 binding affinities in the Karginov data set

Transcript degradation was quantified as the logarithm of the ratio of transcript expression in the lysates of miRNA-transfected and mock-transfected cells. The miRNA-specific EIF2C2 binding was quantified as the ratio of two ratios: EIF2C2-IP of miRNA-transfected and mock-transfected cells and lysates of miRNA-transfected and mock-transfected cells (Supplementary Figure 49).

Extraction of positives and negatives from replicated transfection experiments

Among the transcriptomics data sets we reanalyzed, the experiments performed by Grimson et al. [83], Selbach et al. [191], Baek et al. [8] and Krützfeldt et al. [123] did not feature biological replicates. For these data sets, we considered the top 250 down-regulated (or up-regulated for Krützfeldt et al. [123]) transcripts that carried precisely one 7mer or 8mer seed match. After discarding all seed matches located in the CDS, we ended up with a set of *positive* seed matches. The negatives were obtained through selecting the 250 least-changing transcripts with seed matches, that is the 250 transcripts whose log2 expression fold changes were closest to 0 when comparing the miRNA-transfected samples to

the mock-transfected samples. After discarding all seed matches located in the CDS, we ended up with a set of *negative* seed matches.

The experiments performed by Linsley et al. [140] and Karginov et al. [110] on the other hand featured biological replicates. For these data sets, we applied a method that we designed for selecting transcripts that, with high probability, are affected in expression by the miRNA across all experiments in which the expression of the given miRNA was perturbed (see Supplementary Material in chapter B). Briefly, we first need to calculate, for each pairwise microarray comparison (further referred to as contrast) k, the probability $P_k(f|-)$ that a transcript that is *not* a target, will have a log fold change of f. To estimate the distributions $P_k(f|-)$ we assumed that they are Gaussian with means μ_k and standard deviation σ_k to be estimated from the data for each contrast k. We in addition assumed that transcripts that do not carry at least a heptameric seed-complementary site are unlikely to be real targets, and thus estimated μ_k and σ_k from the observed expression changes of transcripts without such seed matches. We similarly need to calculate, for each contrast k, a distribution $P_k(f|+)$ that a transcript which is a true target of the miRNA, will have a fold-change f. As little is currently known of the distribution of the severity of the effect that miRNAs have on the expression of their targets we assumed as little as possible about the distribution $P_k(f|+)$, namely that a true target must change expression in the right direction, i.e. $f < 0$ for a miRNA over-expression experiment, and $f > 0$ for a miRNA knock-down experiment, and that expression changes are limited to a finite range over which the expression change has a *uniform* distribution. Finally, based on these distributions, we estimate the posterior probability that a transcript with fold change f is a functional target in a given experiment. Details are given in the Supplementary Material (Chapter B). The same procedure was used to construct the sets of positives and negatives from our miR-124 and miR-7 transfection experiments. The process is illustrated in Supplementary Figure 50 and the lists of transcripts with a posterior probability of $\geqslant 0.5$ of being functional in both contrasts of our two miRNA transfection experiments are available for download on Genome Research website. For the negatives we selected those transcripts with minimal sum of squared log2 fold changes in the two experiments. Finally, for the feature analysis, we then proceeded as with experiments where no replicates were performed: we selected 250 positives and 250 negatives according to the criteria defined above and we discarded those cases in which the seed match was in the CDS.

Supplementary Tables 1 and 2 at `http://genome.cshlp.org/content/19/11/2009/suppl/DC1`

Extraction of positives and negatives from ElMMo predictions

From our predictions of miRNA target sites inferred to be under evolutionary selection [70] and for each of the experimentally tested and conserved miRNAs (miR-30a, let-7c, miR-155, miR-1, miR-103, miR-15a, miR-16, miR-106b, miR-20a, miR-141, miR-200a, miR-181a, miR-124 and miR-17), we selected the top 250 target sites in the order of their posterior probability of being under selection. We also selected an equal number of sites least likely to be under selection.

Feature definition and computation

To minimize the ambiguity of attributing a specific response to a miRNA binding site, we only analyzed transcripts that had precisely one miRNA seed match (complementarity to positions 1-7, 2-8, or 1-8 of the miRNA) and the site was at least 100 nucleotides away from either of the boundaries of the 3′ UTR. A sketch of the transcript regions used for the various computations below is shown in Supplementary Figure 51. For each individual putative target site we then computed the following quantities.

SEED ACCESSIBILITY (seed Eopen) was defined in terms of the energy necessary to open the secondary structure of the target in the region binding positions 1-8 of the miRNA. This was computed using the program RNAup of the Vienna package [102] with the following parameters: u=8 (length of the window required to be single-stranded), w=50 (maximal length of the interacting region). The rest of the parameters were left with their default values. Other choices of the w parameter did not qualitatively affect our results (not shown). The negative value of this energy can be viewed as a measure of accessibility.

SITE ACCESSIBILITY (site Eopen) was similarly defined in terms of the energy required to open the secondary structure of the target in a region of 20 nucleotides, anchored at the 3′ end by the seed-complementary region (opposite positions 1-8 of the miRNA). The computation was performed as described above, except that we used a window size u of 20 instead of 8.

ACCESSIBILITY OF THE FLANKS (flanks Eopen) was defined as the average accessibility (defined above) of a window of length 20 contained in the regions of 50 nucleotides upstream or 50 nucleotides downstream of the miRNA target site.

SEED HYBRIDIZATION ENERGY (seed Eduplex) is the energy ΔG_h of the hybrid formed between the seed (position 1-8 of the miRNA) and the seed-complementary site, as given by the RNAduplex program of the Vienna package [102].

MIRNA HYBRIDIZATION ENERGY (target site Eduplex) is the energy of the hybrid formed between the miRNA (positions 1-20) and the miRNA-complementary site, as given by the RNAduplex program.

3′ MIRNA REGION HYBRIDIZATION ENERGY (3′ region Eduplex) is the energy of the hybrid formed between bases 9 to 20 of the miRNA and the 12 nucleotides upstream of the seed complementary region of the mRNA, computed with the RNAduplex program.

PAIRING CONTRIBUTION OF DIFFERENT 3′ REGIONS OF THE MIRNA (pos. 9–12 Eduplex, pos. 13–16 Eduplex, pos. 13–20 Eduplex) is the difference $\Delta G_u - \Delta G_c$ between the minimum binding free energy ΔG_u of the full mRNA-miRNA duplex and the binding free energy ΔG_c of the same duplex under the constraint that the nucleotides 9–12, 13–16 or 13-20 of miRNA are unpaired, respectively. The duplex structure with minimum binding free energy

was computed by the RNAduplex program of the Vienna package. Starting from this structure, we enforced the constraints at positions 9–12, 13–16 and 13–20 and computed the corresponding binding free energy ΔG_c using RNAeval from the Vienna package [102].

SEED INTERACTION ENERGY (seed Einteraction) was defined as $\Delta G = \Delta G_o + \Delta G_h$, where ΔG_o is the energy required to open the secondary structure of the target in the seed-complementary region, and ΔG_h is the energy of the hybrid formed between the seed and the seed-complementary site. ΔG_o is obtained as described in the paragraph "Seed accessibility" above, and ΔG_h is computed using the RNAduplex program [102] with default parameters. Note that we neglected the energy possibly required to open the structure of the seed region of the miRNA. The probability of interaction with the seed region of the miRNA (seed Pinteraction) is the corresponding probability, as computed by RNAup.

MIRNA INTERACTION ENERGY (target site Einteraction) was similarly defined as $\Delta G = \Delta G_o + \Delta G_h$, where ΔG_o is the energy required to open the secondary structure of the target in the miRNA-binding region of 20 nucleotides anchored at the seed (as described above), and ΔG_h is the energy of the hybrid formed between the miRNA and the miRNA-complementary site. ΔG_o is obtained as described at point 2 above, and ΔG_h is computed using the RNAduplex program [102] with default parameters.

FLANKS A, C, G AND U CONTENTS were defined as the proportions of A, C, G and U nucleotides within 50 nucleotides upstream and 50 nucleotides downstream of the miRNA binding site of 20 nucleotides, anchored downstream by the seed-matching region.

3'UTR A, C, G AND U CONTENTS were defined as the proportions of A, C, G and U nucleotides within the 3'UTR harboring the miRNA binding site.

TRANSCRIPT A, C, G AND U, AU CONTENTS were defined as the proportions of A, C, G, U and A+U nucleotides in the transcript harboring the miRNA binding site.

TRANSCRIPT AND 3'UTR LENGTH were obtained from the RefSeq sequence and annotation.

RELATIVE POSITION was computed by dividing the position in the 3'UTR marking the beginning of the seed complementary region by the 3'UTR length.

RELATIVE DISTANCE TO 3'UTR BOUNDARY was computed similarly, dividing the minimal distance from the beginning of the seed complementary region to the STOP codon or the poly-A tail by the length of the 3'UTR.

FLANKING SEQUENCE OPTIMIZATION was designed to measure the extent to which the nucleotide composition of the regions flanking a miRNA binding site explains the accessibility of the miRNA binding site. For each target site we generated 100 variants in which we randomized, independently of each other, the sequence

of the 50 nucleotides upstream and of the 50 nucleotides downstream of the miRNA target site, while keeping the mono-nucleotide frequencies in these regions constant. For the randomized variants we recomputed the accessibility of the miRNA binding site as described above. We then calculated the z-statistic of the real sequence relative to the randomized variants. This computation gave us one set of z-statistics for the positives and one for the negatives. We finally used the t-test to compare the means of the two distributions of z-statistics.

ELMMO is the posterior probability that a seed complementary region is under evolutionary selective pressure described in Gaidatzis et al. [70].

Testing different linear models for predicting various types of miRNA target sites

We divided all the data sets that we studied here into three groups, as described in section 3.2.3. We then performed a principal component analysis to determine a set of 14 non-redundant features (listed in the legend of Figure 9). We then used these features to train three independent generalized linear models (GLM) with logit link function [150] on the transcriptomics data sets, on the proteomics data sets, and a mixture of the transcriptomics and proteomics data sets. In the latter case, we weighted each putative miRNA target site proportionally to the inverse of the data set size, to have the resulting model minimize the prediction error equally on both data sets.

To avoid over-estimating the performance of the three GLMs when testing them on the data sets on which they were trained, we performed 10-fold cross-validation. In other words, we split our data set in 10 parts, trained the model using the first 9 parts of the data set, and evaluated its sensitivity and specificity on the last. We reiterated this procedure 10 times and used the numbers that came out of it to plot the cross-validated receiver operating characteristic (ROC) curve. The ROC curves for GLMs trained on a different data sets and for other miRNA target predictions algorithms were computed using the standard procedure [201].

To simulate ROC curves from biological replicates of varying reproducibilities, we sampled 25000 points from bivariate gaussians with correlation coefficients r of 0.4 and 0.8, which covers the range of reproducibilities of log2 fold changes that we observed in the experimental data sets that we analyzed here. We then considered the 10% (2500) smallest values from the first simulated replicate as fold changes in a transfection experiment for "true target sites" and attempted to use the second simulated replicate to predict the "true target sites". The two ROC curves show to what extent knowing one simulated data set enables one to predict the other depending on whether the replicates are in moderate ($r = 0.4$) or good agreement with each other ($r = 0.8$).

Evaluating the competition between the endogenous and the transfected miRNA

Khan et al. [118] recently reported that transfected miRNAs compete with the endogenous miRNAs for RISC loading. To evaluate this effect in the context of our study, we applied the analysis methods of Khan

et al. [118] to all 44 microarrays performed in the HCT116 Dicer -/-, 8, 10, 14 and 24 hours after miRNA or siRNA transfection (GEO accessions: GSM156513, GSM156514, GSM156515, GSM156516, GSM156517, GSM156518, GSM156519, GSM156520, GSM156567, GSM156568, GSM156569, GSM156570, GSM156571, GSM156572, GSM156573, GSM156574, GSM156525, GSM156526, GSM156527, GSM156536, GSM156521, GSM156522, GSM156523, GSM156524, GSM156531, GSM156532, GSM156533, GSM156534, GSM156545, GSM156546, GSM156547, GSM156548, GSM156553, GSM156554, GSM156557, GSM156559, GSM156565, GSM156566, GSM156575, GSM156576, GSM156577, GSM156578, GSM156579, GSM156580, GSM156581) and to microarrays that monitored the mRNA expression changes at 8 and 32 hours after the transfection of five miRNAs published by Selbach et al. [191].

To be able to compare our results with those of Khan et al. [118], we modified slightly the microarray data processing described in sections 3.4 and 3.4: at the step where we choose a representative RefSeq mRNA for each gene monitored on the microarray we chose the RefSeq mRNA with longest 3′ UTR instead of the RefSeq with median length 3′ UTR.

We determined the set "X" of mRNAs whose 3′ UTRs carried a match to positions 2 to 8 of the transfected miRNA. We then determined the set "D" of mRNAs carrying a 2-8 seed match to one of the top 10 miRNA families most expressed in the cell line (HCT116 Dicer -/- or HeLa) used in the experiment. We used the miRNA family expression profiles reported on Supplementary Figure 2 of Khan et al. [118]. We determined the set "B" of mRNAs that carried seed matches to neither the transfected miRNA nor the top 10 endogenous miRNA families. Finally, we applied a linear transformation to the log2 fold change such that the log fold changes of the mRNA belonging to the B set had mean 0 and variance 1.

We then computed the average log fold changes of the X, $X \cap D$, $X \setminus D$, $D \setminus X$ and B mRNA sets. Doing so for each set of mRNAs and for each time point gave us one measurement of mRNA log fold change per experiment (*i.e.* per transfected miRNA), which we combined by averaging over all experiments performed at the same time point and computing the 95% confidence interval on the mean log fold changes.

3.5 ACKNOWLEDGMENTS

We thank Thomas Tuschl for support to perform the experiments, suggestions and discussions. We are grateful to Wenxiang Zhang and Connie Zhao (Rockefeller University Genomics Resource Center) for mRNA array analyses, to Erik van Nimwegen (Biozentrum, University of Basel) for suggesting the model to identify functional sites from multiple experiments, and to Ted Karginov for providing us the raw Agilent data of the miR-124 EIF2C-IP. J.H. was supported by the Swiss National Fund Grant #3100A0-114001 to Mihaela Zavolan. M.L. was partially supported by an Irma T. Hirschl Postdoctoral Fellowship. In addition M.L. was supported by NIH grant GM068476 to Thomas Tuschl. We also thank Lukas Burger, Erik van Nimwegen and Walter Keller (Biozentrum, University of Basel) for critical comments on the manuscript.

4

TRANSCRIPTOME-WIDE IDENTIFICATION OF RNA-BINDING PROTEIN AND MICRORNA TARGET SITES BY PAR-CLIP

The method presented in this chapter was developped in collaboration with the Tuschl lab at Rockefeller University, New York and were originally published in Cell [90]

4.1 INTRODUCTION

Gene expression in eukaryotes is extensively controlled at the post-transcriptional level by hundreds of miRNAs, which are bound by Argonaute (Ago/EIF2C) proteins and mediate destabilization and/or inhibition of translation of partially complementary target mRNAs [13]. But Ago is just one out of hundreds of RNA-binding proteins (RBPs) and ribonucleoprotein complexes (RNPs) [151] that modulate the maturation, stability, transport, editing and translation of RNA transcripts in vertebrates [147, 156, 200]. Each of these RBPs contain one or more domains able to specifically recognize target transcripts. To understand how the interplay of these RNA-binding factors affects the regulation of individual transcripts, high resolution maps of in vivo protein-RNA interactions are necessary [116].

A combination of genetic, biochemical and computational approaches is typically applied to identify RNA-RBP or RNA-RNP interactions. Microarray profiling of RNAs associated with immunopurified RBPs (RIP-Chip) [209] defines targets at a transcriptome level, but its application is limited to the characterization of kinetically stable interactions and does not directly identify the RBP recognition element (RRE) within the long target RNA. Nevertheless, RREs with higher information content can be derived computationally from RIP-Chip data, e.g. for HuR [43] or for Pumilio [76].

More direct RBP target site information is obtained by combining in vivo UV crosslinking [81, 222] with immunoprecipitation [49, 149] followed by the isolation of crosslinked RNA segments and cDNA sequencing (CLIP) [214]. CLIP was used to identify targets of the splicing regulators NOVA1 [136], FOX2 [235] and SFRS1 [186] as well as U3 snoRNA and pre-rRNA [80], pri-miRNA targets for HNRNPA1 [87], EIF2C2/AGO2 protein binding sites [38] and ALG-1 target sites in C. elegans [241]. CLIP is limited by the low efficiency of UV 254 nm RNA-protein crosslinking, and the location of the crosslink is not readily identifiable within the sequenced crosslinked fragments, raising the question of how to separate UV-crosslinked target RNA segments from background non-crosslinked RNA fragments also present in the sample.

Here we describe an improved method for isolation of segments of RNA bound by RBPs or RNPs, referred to as PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation). To facilitate crosslinking, we incorporated 4-thiouridine (4SU) into transcripts of cultured cells and identified precisely the RBP binding sites by scoring for thymidine (T) to cytidine (C) transitions in the sequenced cDNA. We uncovered tens of thousands of binding sites for several important RBPs and RNPs and assessed the regulatory impact of binding on their targets. These findings underscore the complexity of post-transcriptional regulation of cellular systems.

4.2 RESULTS

4.2.1 *Photoactivatable nucleosides facilitate RNA-RBP crosslinking in cultured cells*

Random or site-specific incorporation of photoactivatable nucleoside analogs into RNA in vitro has been used to probe RBP- and RNP-RNA interactions [121, 152]. Several of these photoactivatable nucleosides are readily taken up by cells without apparent toxicity and have been used for in vivo crosslinking [65]. We applied a subset of these nucleoside analogs (Figure 11A) to cultured cells expressing the FLAG/HA-tagged RBP IGF2BP1 followed by UV 365 nm irradiation. The crosslinked RNA-protein complexes were isolated by immunoprecipitation, and the covalently bound RNA was partially digested with RNase T1 and radiolabeled. Separation of the radiolabeled RNPs by SDS-PAGE indicated that 4SU-containing RNA crosslinked most efficiently to IGF2BP1. Compared to conventional UV 254 nm crosslinking, the photoactivatable nucleosides improved RNA recovery 100- to 1000-fold, using the same amount of radiation energy (Figure 11B). We refer to our method as PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) (Figure 11C).

We evaluated the cytotoxic effects upon exposure of HEK293 cells to 100 μM and 1 mM of 4SU or 6SG in tissue culture medium over a period of 12 h by mRNA microarrays. The mRNA profiles of 4SU or 6SG treated cells were very similar to those of untreated cells (Table S1), suggesting that the conditions for endogenous labeling of transcripts were not toxic.

To guide the development of bioinformatic methods for identification of binding sites, we first studied human Pumilio 2 (PUM2), a member of the Puf-protein family (Figure 12A) known for its highly sequence-specific RNA binding [224].

4.2.2 *Identification of PUM2 mRNA targets and its RRE*

PUM2 protein crosslinked well to 4SU-labeled cellular transcripts (Figure 12B). The crosslinked segments were converted into a cDNA library and Solexa sequenced [89]. The sequence reads were aligned against the human genome and EST databases. Reads mapping uniquely to the genome with up to one mismatch, insertion or deletion were used to build clusters of sequence reads (Figure 12C, Supplementary Methods, and Table S2). We obtained 7,523 clusters originating from about 3,000 unique transcripts, 93% of which were found within the 3′ untranslated region (UTR) (Figure 53) in agreement with previous studies [230]. All sequence clusters with mapping and annotation information are available online[1].

PhyloGibbs analysis [197] of the top 100 most abundantly sequenced clusters (Table S3), as expected, yielded the PUM2 RRE, UGUANAUA [73] (Figure 12D). Unexpectedly, over 70% of all sequence reads that gave rise to clusters showed a T to C mutation compared to the genome (Figure 53). Ranking of sequence read clusters according to the frequency of T to C mutation further enriched for the PUM2 RRE (Figure 53) indicating that the T to C mutation is diagnostic of sequences interacting with the RBP. The T to C changes were not randomly distributed: the

1 http://www.mirz.unibas.ch/restricted/clipdata/RESULTS/index.html

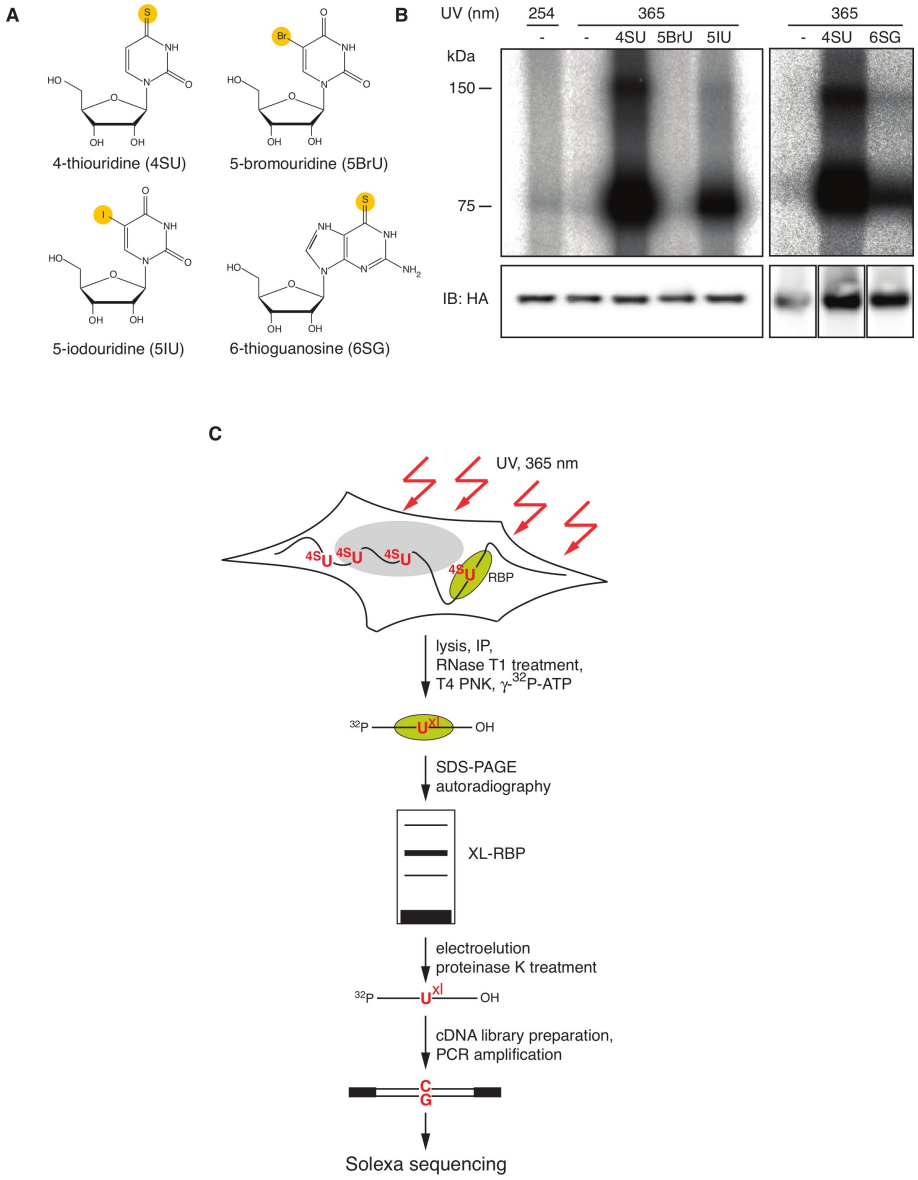


Figure 11: PAR-CLIP methodology (A) Structure of photoactivatable nucleosides (B) Phosphorimages of SDS-gels that resolved 5′-32P-labeled RNA-FLAG/HA-IGF2BP1 immunoprecipitates (IPs) prepared from lysates from cells that were cultured in media in the absence or presence of 100 µM photoactivatable nucleoside and crosslinked with UV 365 nm. For comparison, a sample prepared from cells crosslinked with UV 254 nm, was included. Lower panels show immunoblots probed with an anti-HA antibody. (C) Illustration of PAR-CLIP. 4SU-labeled transcripts were crosslinked to RBPs and partially RNase-digested RNA-protein complexes were immunopurified and size-fractionated. RNA molecules were recovered and converted into a cDNA library and deep sequenced.

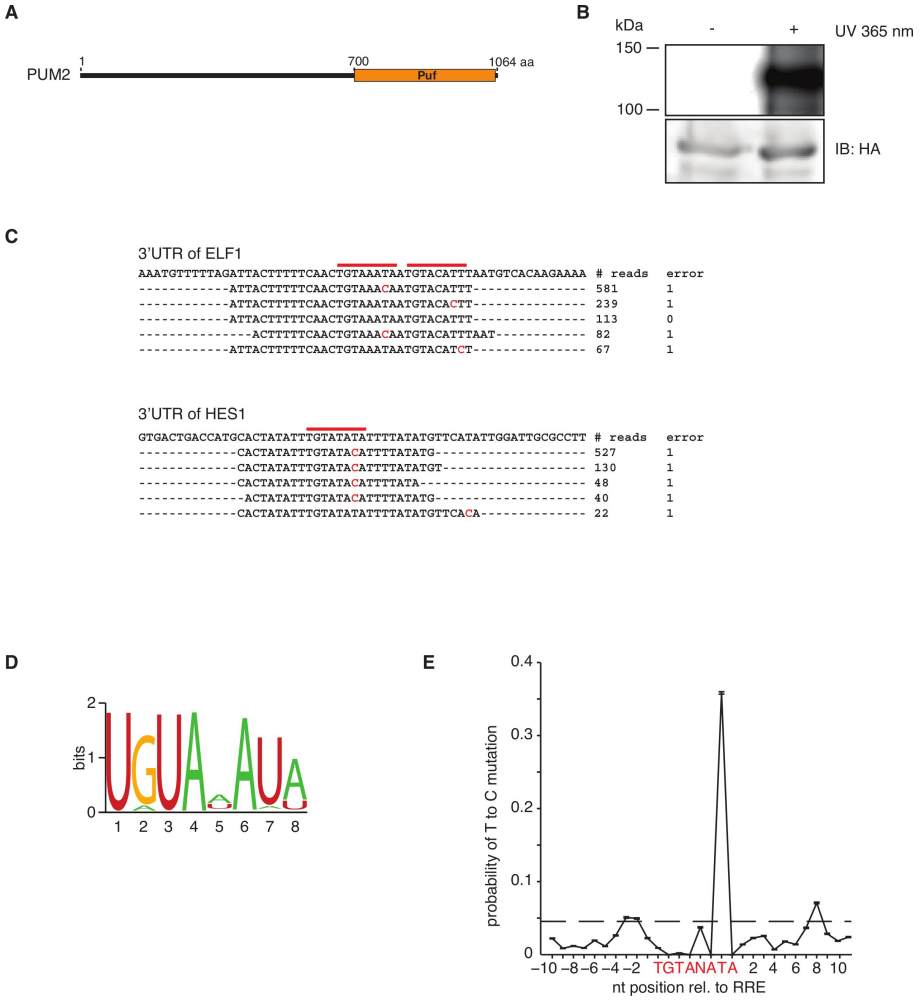


Figure 12: RNA recognition by PUM2 protein (A) Domain structure of PUM2 protein. (B) Phosphorimage of SDS-gel of radiolabeled FLAG/HA-PUM2-RNA complexes from non-irradiated or UV-irradiated 4SU-labeled cells. The lower panel shows an anti-HA immunoblot. (C) Alignments of PAR-CLIP cDNA sequence reads to corresponding regions in the 3'UTR of ELF1 and HES1 Refseq transcripts. The number of sequence reads (# reads) and mismatches (errors) are indicated. Red bars indicate the PUM2 recognition motif and red-letter nucleotides indicate T to C sequence changes. (D) Sequence logo of the PUM2 recognition motif generated by PhyloGibbs analysis of the top 100 sequence read clusters. (E) T to C positional mutation frequency for PAR-CLIP clusters anchored at the 8-nt recognition motif from all motif-containing clusters (Table S3). The dashed line represents the average T to C mutation frequency within these clusters. See also Figure 53.

T corresponding to U7 of the RRE mutated at higher frequency compared to the Ts corresponding to U1 and U3 (Figure 12E). Our analyses suggest that the reverse transcriptase specifically misincorporated dG across from crosslinked 4SU residues and that local amino acid environment also affected crosslinking efficiency. Uridines proximal to the RRE also exhibited an increased T to C mutation frequency, indicating that crosslinks also form in close proximity to an RRE and that our method even captured PUM2 binding sites that did not have a U7 in its RRE.

4.2.3 *Identification of QKI RNA targets and its RRE*

To further validate our method, we applied it to the RBP Quaking (QKI), which contains a single heterogeneous nuclear ribonucleoprotein K homology (KH) domain (Figures 13A,B). The RRE ACUAAY was determined by SELEX [71], but in vivo targets are largely undefined. Mice with reduced expression of QKI show dysmyelination and develop rapid tremors or "quaking" 10 days after birth. Previous studies suggested that QKI participates in pre-mRNA splicing, mRNA export, mRNA stability and protein translation [36].

PhyloGibbs analysis of the 100 most abundantly sequenced clusters (Table S3) yielded the RRE AYUAAY (Figures 13C,D), similar to a motif identified by SELEX [71]. We found approx. 6,000 clusters mapping to 2,500 transcripts. Close to 75% of these clusters were derived from intronic sequences, supporting the hypothesis that QKI is a splicing regulator (Chenard and Richard, 2008) and 70% of the remaining exonic clusters fall into 3'UTRs (Figure 54).

Mutation analysis of the clustered sequence reads showed that the T corresponding to U2 in AUUAAY was frequently altered to C whereas the T corresponding to U3 in AUUAAY or ACUAAY remained unaltered (Figure 13E). Crosslinking of 4SU residues located in immediate vicinity to the RRE was mostly responsible for exposing the motif with C2, showing that crosslinking inside the recognition element is not a precondition for its identification. Hence, the discovery of RREs is unlikely to be prevented by sequence-dependent crosslinking biases as long as deep enough sequencing captures these interaction sites at and nearby the RRE.

4.2.4 *T to C mutations occur at the crosslinking sites*

To better characterize the T to C transition observed in crosslinked RNA segments, we UV 365 nm crosslinked oligoribonucleotides containing single 4SU substitutions to recombinant QKI (Figures 13F,G). The crosslinking efficiency varied 50-fold and mirrored the results of the mutational analysis (Figure 13G). The least effective crosslinking was observed for placement of 4SU at position 3 of the QKI RRE (4SU9), and the most effective crosslinking was found at position 2 of the QKI RRE (4SU10); the crosslinking efficiency for two positions outside of the RRE (4SU2 and 4SU4) was intermediate. Neither of these substitutions affected RNA-binding to recombinant QKI protein as determined by gel-shift analysis, whereas mutations of the recognition element weakened the binding between 2.5- and 9-fold (Table S1).

Next, we sequenced libraries prepared from non-crosslinked as well as QKI-protein-crosslinked oligoribonucleotides containing 4SU at in-

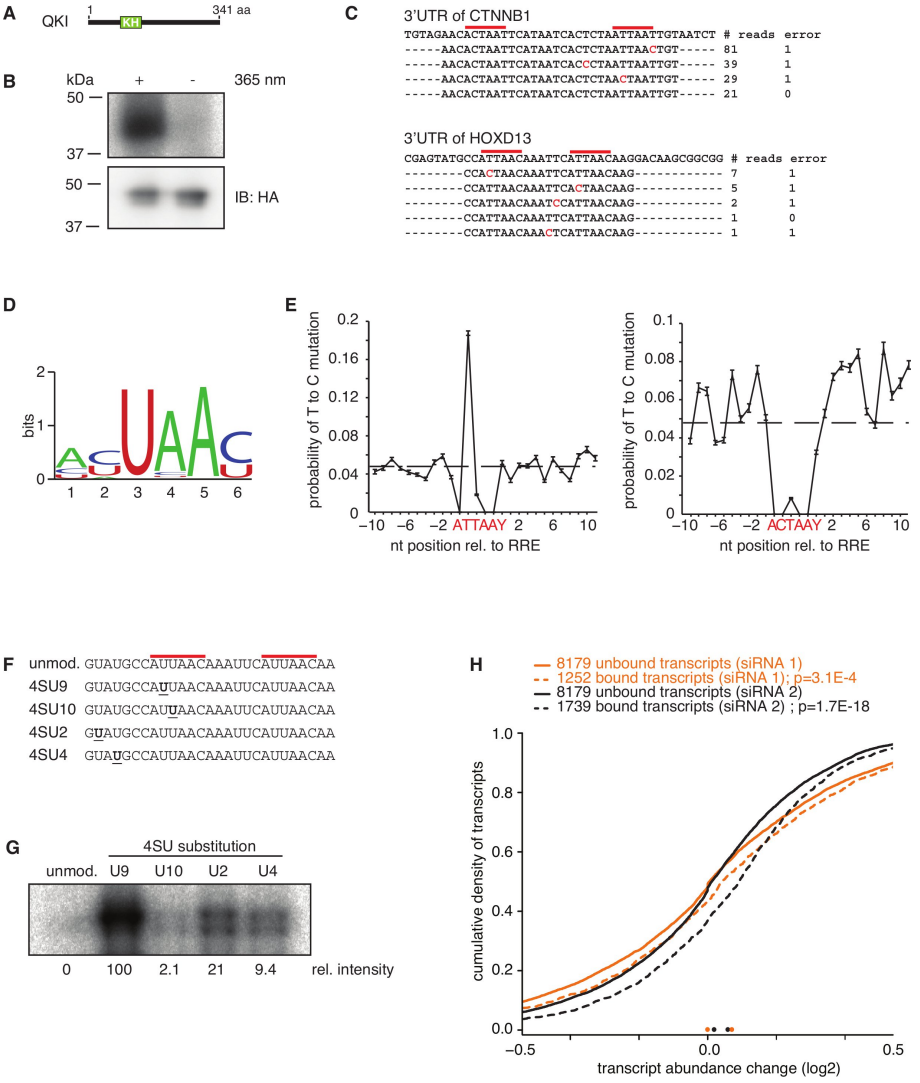


Figure 13: RNA recognition by QKI protein (A) Domain structure of QKI protein (B) Phosphorimage of SDS-gel resolving radiolabeled RNA crosslinked to FLAG/HA-QKI IPs from non-irradiated or UV-irradiated 4SU-labeled cells. The lower panel shows the anti-HA immunoblot. (C) Alignments of PAR-CLIP cDNA sequence reads to the corresponding regions in the 3'UTRs of the CTNNB1 and HOXD13 transcripts. Red bars indicate the QKI recognition motif and red-letter nucleotides indicate T to C sequence changes. (D) Sequence logo of the QKI recognition motif generated by PhyloGibbs analysis of the top 100 sequence read clusters. (E) T to C positional mutation frequency for PAR-CLIP clusters anchored at the AUUAAY (left panel) and ACUAAY (right panel) RRE (Table S3); Y = U or C. The dashed line represents the average T to C mutation frequency within these clusters. (F) Sequences of synthetic 4SU-labeled oligoribonucleotides with QKI recognition motifs, derived from a sequence read cluster aligning to the 3'UTR of HOXD13 shown in (C) 4SU-modified residues are underlined. (G) Phosphorimage of SDS-gel resolving recombinant QKI protein after crosslinking to radiolabeled synthetic oligoribonucleotides shown in (F). (H) Stabilization of QKI-bound transcripts upon siRNA knockdown. Changes in mRNA levels upon QKI knockdown by two distinct siRNAs were measured by microarray analysis. Shown are the distributions of changes upon siRNA transfection for transcripts that did (dashed lines) or did not (solid lines) contain QKI PAR-CLIP clusters. See also Figure 54.

dicated positions (Figure 13F). The fraction of sequence reads with T to C changes obtained from non-irradiated 4SU-containing oligoribonucleotides varied between 10 and 20%, and increased to 50 to 80% upon crosslinking (Table S1). The variation of the degree of T to C changes in the crosslinked samples is most likely determined by background of non-crosslinked oligoribonucleotides. Presumably, the T to C transition frequency is increased upon crosslinking as a direct consequence of a chemical structure change of the 4SU nucleobase upon crosslinking to protein amino acid side chains, resulting in altered stacking or hydrogen bond donor/acceptor properties directing the preferential incorporation of dG rather than dA during reverse transcription (Figure 53). At the doses of 4SU applied to cultured cells, about 1 out of 40 uridines was substituted by 4SU as determined by HPLC analysis of the nucleoside composition of total RNA. Assuming a 20% T to C conversion rate for a non-crosslinked 4SU-labeled site, we estimated that the average T to C conversion rate of 40-nt sequence reads derived from background non-crosslinked sequences will be near 5%. Clusters of sequence reads with average T to C conversion above this threshold, irrespective of the number of sequence reads, most certainly represent crosslinking sites. The ability to separate signal from noise by focusing on clusters with a high frequency of T to C mutations rather than clusters with the largest number of reads, represents a major enhancement of our method over UV 254 nm crosslinking methods.

To assess whether the transcripts identified by PAR-CLIP are regulated by QKI, we analyzed the mRNA levels of mock-transfected and QKI-specific siRNA-transfected cells with microarrays. Transcripts crosslinked to QKI were significantly upregulated upon siRNA transfection, indicating that QKI negatively regulates bound mRNAs (Figure 13H), consistent with previous reports of QKI being a repressor [36].

4.2.5 *Identification of IGF2BP family RNA targets and its RRE*

We then applied PAR-CLIP to the FLAG/HA-tagged insulin-like growth factor 2 mRNA-binding proteins 1, 2, and 3 (IGF2BP1-3) (Figures 14A,B), a family of highly conserved proteins that play a role in cell polarity and cell proliferation [237]. These proteins are predominantly expressed in the embryo and regulate mRNA stability, transport and translation. They are re-expressed in various cancers [22, 45] and IGF2BP2 has been associated with type-2 diabetes [187]. The IGF2BPs are highly similar and contain six canonical RNA-binding domains, two RNA recognition motifs (RRMs) and four KH domains (Figure 14A). Therefore, target recognition for this protein family appears complex, with only a small number of coding and non-coding RNA targets being known so far. A precise definition of the RREs is missing [237].

The three IGF2BPs recognized a highly similar set of target transcripts (Table S1), suggesting similar and redundant functions. PhyloGibbs analysis of the clusters derived from mRNAs (Figure 14C and Table S3) yielded the sequence CAUH (H=A, U, or C) as the only consensus recognition element (Figure 14D), contained in more than 75% of the top 1000 clusters for IGF2BP1, 2 or 3 (Figure 55). In total, we identified over 100,000 sequence clusters recognized by the IGF2BP family that map to about 8,400 protein-coding transcripts. The annotation of the clusters was predominantly exonic (ca. 90%) with a slight preference for 3'UTR relative to coding sequence (CDS) (Figure 55).

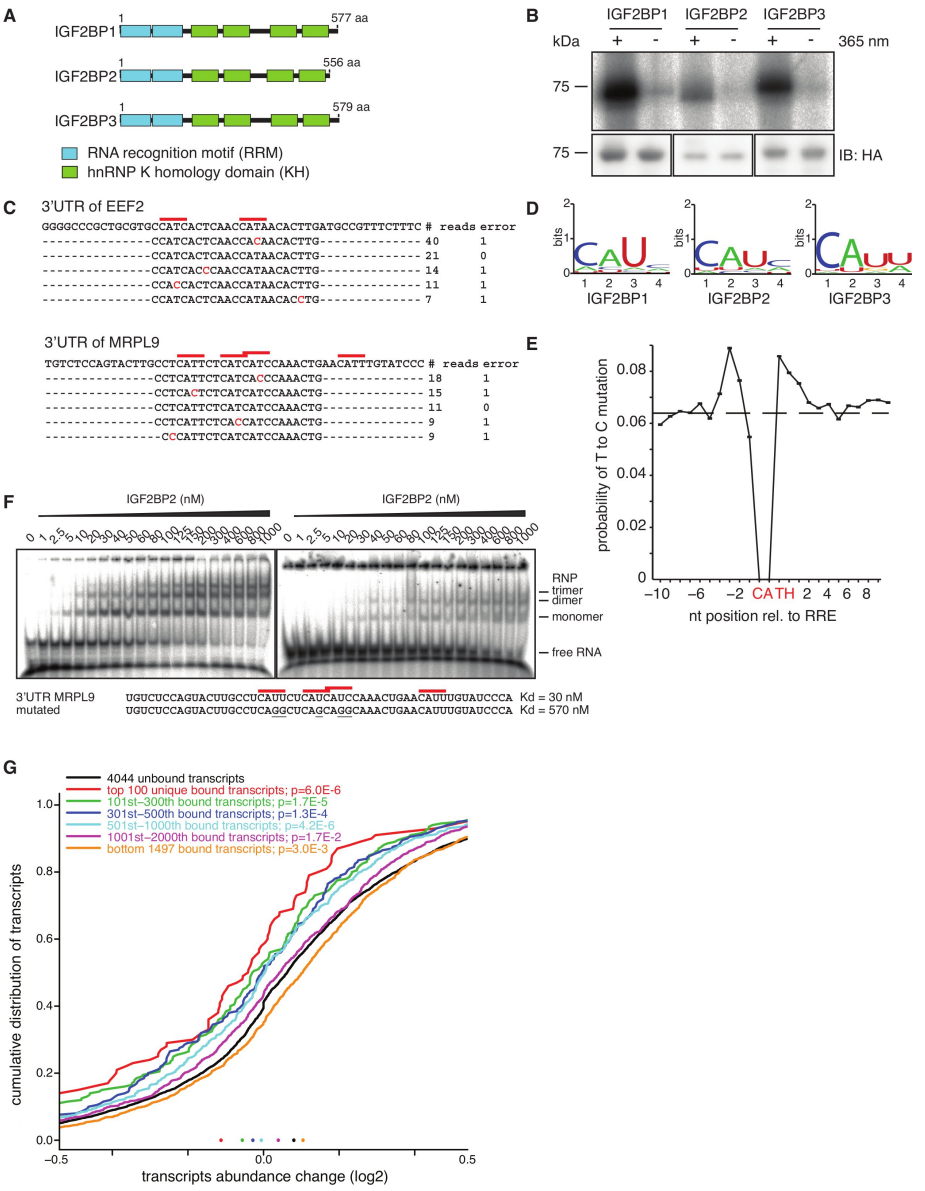


Figure 14: RNA recognition by the IGF2BP protein family (A) Domain structure of IGF2BP1-3 proteins. (B) Phosphorimage of an SDS-gel resolving radiolabeled RNA crosslinked to FLAG/HA-IGF2BP1-3 IPs. The lower panel shows anti-HA immunoblots. (C) Alignments of IGF2BP1 PAR-CLIP cDNA sequence reads to the corresponding regions of the 3'UTRs of EEF2 and MRPL9 transcripts. Red bars indicate the 4-nt IGF2BP1 recognition motif and nucleotides marked in red indicate T to C sequence changes. (D) Sequence logo of the IGF2BP1-3 RRE generated by PhyloGibbs analysis of the top 100 sequence read clusters. (E) T to C positional mutation frequency for PAR-CLIP clusters anchored at the 4-nt recognition motif from all motif-containing clusters (Table S3). The dashed line represents the average T to C mutation frequency within these clusters. (F) Phosphorimage of native PAGE resolving complexes of recombinant IGF2BP2 protein with wild-type (left panel) and mutated target oligoribonucleotide (right panel). Sequences and dissociation constants (Kd) are indicated. (G) Destabilization of IGF2BP-bound transcripts upon siRNA knockdown of IGF2BP1, 2, and 3. Distributions of transcript level changes for IGF2BP1-3 PAR-CLIP target transcripts versus non-targeted transcripts are shown. See also Figures 55 and 56.

The mutation frequency of all sequence tags containing the element CAUH (H = A, C, or U) showed that the crosslinked residue was positioned inside the motif, or in the immediate vicinity (Figure 14E). The consensus motif CAUH was found in more than 75% of the top 1000 targeted transcripts, followed in more than 30% by a second motif, predominantly within a distance of three to five nucleotides (Figure 55). In vitro binding assays showed that nucleotide changes of the CAUH motif decreased, but did not abolish the binding affinity (Figure 14F and Table S1).

To test the influence of IGF2BPs on the stability of their interacting mRNAs, as reported previously for some targets [237], we simultaneously depleted all three IGF2BP family members using siRNAs and compared the cellular RNA from knockdown and mock-transfected cells on microarrays. The levels of transcripts identified by PAR-CLIP decreased in IGF2BP-depleted cells, indicating that IGF2BP proteins stabilize their target mRNAs. Moreover, transcripts that yielded clusters with the highest T to C mutation frequency were most destabilized (Figure 14G), indicating that the ranking criterion that we derived based on the analysis of PUM2 and QKI data generalizes to other RBPs.

For comparison to conventional and high-throughput sequencing CLIP [136, 214], we also sequenced cDNA libraries prepared from UV 254 nm crosslinking. Of the 8,226 clusters identified by UV 254 nm crosslinking of IGF2BP1, 4,795 were found in the PAR-CLIP dataset. Although UV 254 nm crosslinking identified the identical segments of a target RNA as PAR-CLIP, the position of the crosslink could not be readily deduced, because no abundant diagnostic mutation was observed (Figure 56).

4.2.6 *Identification of miRNA targets by AGO and TNRC6 family PAR-CLIP*

To test our approach on RNP complexes, we selected the protein components mediating miRNA-guided target RNA recognition. In animal cells, miRNAs recognize their target mRNAs through base-pairing interactions involving mostly 6-8 nucleotides at the 5' end of the miRNA (the so called "seed") [13]. Target sites were thought to be predominantly located in the 3'UTRs of mRNAs, and computational miRNA target prediction methods frequently resort to identification of evolutionarily conserved sites that are located in 3'UTRs and are complementary to miRNA seed regions [13, 175]. We isolated mRNA fragments bound by miRNPs from HEK293 cell lines stably expressing FLAG/HA-tagged AGO or TNRC6 family proteins [130]. The AGO IPs revealed two prominent RNA-crosslinked bands of 100 and 200 kDa, representing AGO, and likely TNRC6 and/or DICER1 protein. The TNRC6 IPs showed one prominent RNA-crosslinked protein of 200 kDa (Figure 15A).

From clusters (Figure 15B) formed by at least 5 PAR-CLIP sequence reads and containing more than 20% T to C transitions (Table S2), we extracted 41 nt long regions centered over the predominant T to C transition or crosslinking site. The length of the crosslink-centered regions (CCRs) was selected to include all possible registers of miRNA/target-RNA pairing interactions relative to the crosslinking site.

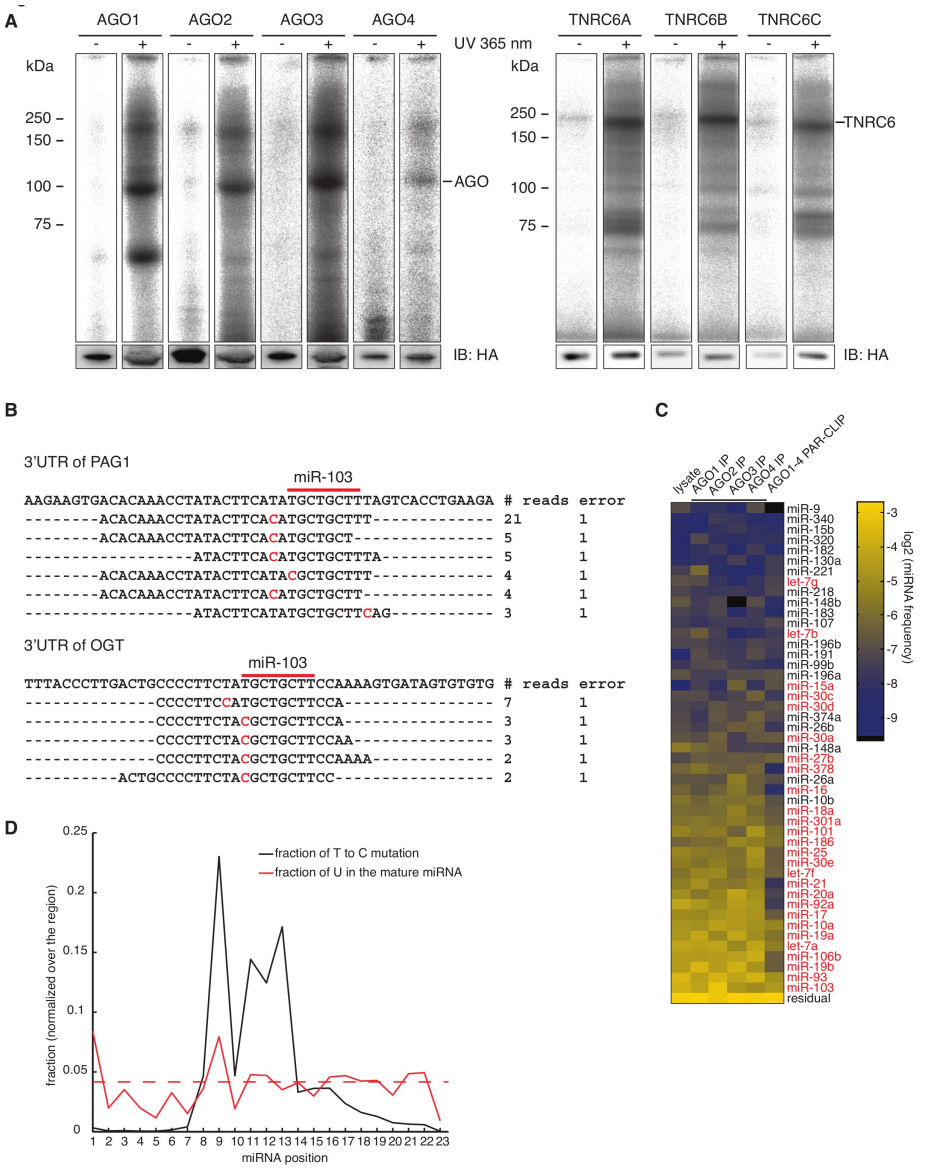


Figure 15: AGO protein family and TNRC6 family PAR-CLIP (A) Phosphorimage of SDS-gels resolving radiolabeled RNA crosslinked to the FLAG/HA-AGO1-4 and FLAG/HA-TNRC6A-C IPs. The lower panel shows the immunoblot with an anti-HA antibody. (B) Alignment of AGO PAR-CLIP cDNA sequence reads to the corresponding regions of the 3'UTRs of PAG1 and OGT. Red bars indicate the 8-nt miR-103 seed complementary sequence and nucleotides marked in red indicate T to C mutations. (C) miRNA profiles from RNA isolated from untreated HEK293 cells, non-crosslinked FLAG/HA-AGO1-4 IPs, and combined AGO1-4 PAR-CLIP libraries. The color code represents relative frequencies determined by sequencing. miRNAs indicated in red were inhibited by antisense oligonucleotides for the transcriptome-wide characterization of the destabilization effect of miRNA binding. (D) T to C positional mutation frequency for miRNA sequence reads is shown in black, and the normalized frequency of occurrence of uridines within miRNAs is shown in red. The dashed red line represents the normalized mean U frequency in miRNAs. See also Figure 57.

PAR-CLIP of individual AGO proteins yielded on average about 4,000 clusters that overlapped, supporting our earlier observation that AGO1-4 bound similar sets of transcripts [130]. We therefore combined the sequence reads obtained from all AGO experiments, which yielded 17,319 clusters of sequence reads at a cut-off of 5 reads (Table S4). These clusters distributed across 4,647 transcripts with defined GeneIDs, corresponding to 21% of the 22,466 unique HEK293 transcripts that we identified by digital gene expression (DGE).

PAR-CLIP of individual TNRC6 proteins yielded on average about 600 clusters that also overlapped substantially, again consistent with our observation that TNRC6 family proteins bind similar transcripts [130]. We therefore combined all sequence reads from all TNRC6 experiments, yielding 1,865 clusters and CCRs (Table S4). More than 50% of these TNRC6 CCRs fell within 25 nt of an AGO CCR, and 26% overlapped by at least 75%, indicating that AGO and TNRC6 members bind to the same sites (Figure 57).

4.2.7 *Comparison of miRNA profiles from AGO PAR-CLIP to non-crosslinked miRNA profiles*

To relate the potential miRNA-target-siteâcontaining CCRs to the endogenously expressed miRNAs, we determined the miRNA profiles from total RNA isolated from HEK293 cells, and miRNAs isolated from non-crosslinked AGO1-4 IPs by Solexa sequencing [89], and compared them to the profile from the miRNAs present in the combined AGO1-4 PAR-CLIP library. miRNA profiles obtained from total RNA and IP of the four AGO proteins in non-crosslinked cells correlated well (Figure 15C and Table S5) supporting our observation that AGO1-4 bind the same targets [130]. The most abundant among the 557 identified miRNAs and miRNAs* were miR-103 (7% of miRNA sequence reads), miR-93 (6.5%), and miR-19b (5.5%). The 25 and 100 most abundant miRNAs accounted for 72% and 95% of the total of miRNA sequence reads, respectively. Comparison of the miRNA profile derived from the combined AGO PAR-CLIP library with the combined non-crosslinked libraries showed a good correlation (Spearman correlation coefficient of 0.56, Figure 15C and Figure 57A).

Importantly, in the AGO PAR-CLIP library, the majority of miRNA sequence reads derived from prototypical miRNAs [129] displayed T to C conversion near or above 50%. The T to C conversion was predominantly concentrated within positions 8 to 13 (Figure 15D), residing in the unpaired regions of the AGO protein ternary complex [227]. Five of the 100 most abundant miRNAs in HEK293 cells lack uridines at position 8-13, yet only 2 of those miRNAs, miR-374a and b, showed no crosslinking, because uridines at residues 14 and higher can still be crosslinked (Table S5). This frequency of crosslinks was substantially lower in the miRNAs whose expression did not correlate between AGO-IP and AGO PAR-CLIP samples compared to the miRNAs whose expression correlated well (Figure 57).

4.2.8 *mRNAs interacting with AGOs contain miRNA seed complementary sequences*

Independent of any pairing models for miRNAs and their targets, we first determined the enrichment of all 16,384 possible 7-mers within the

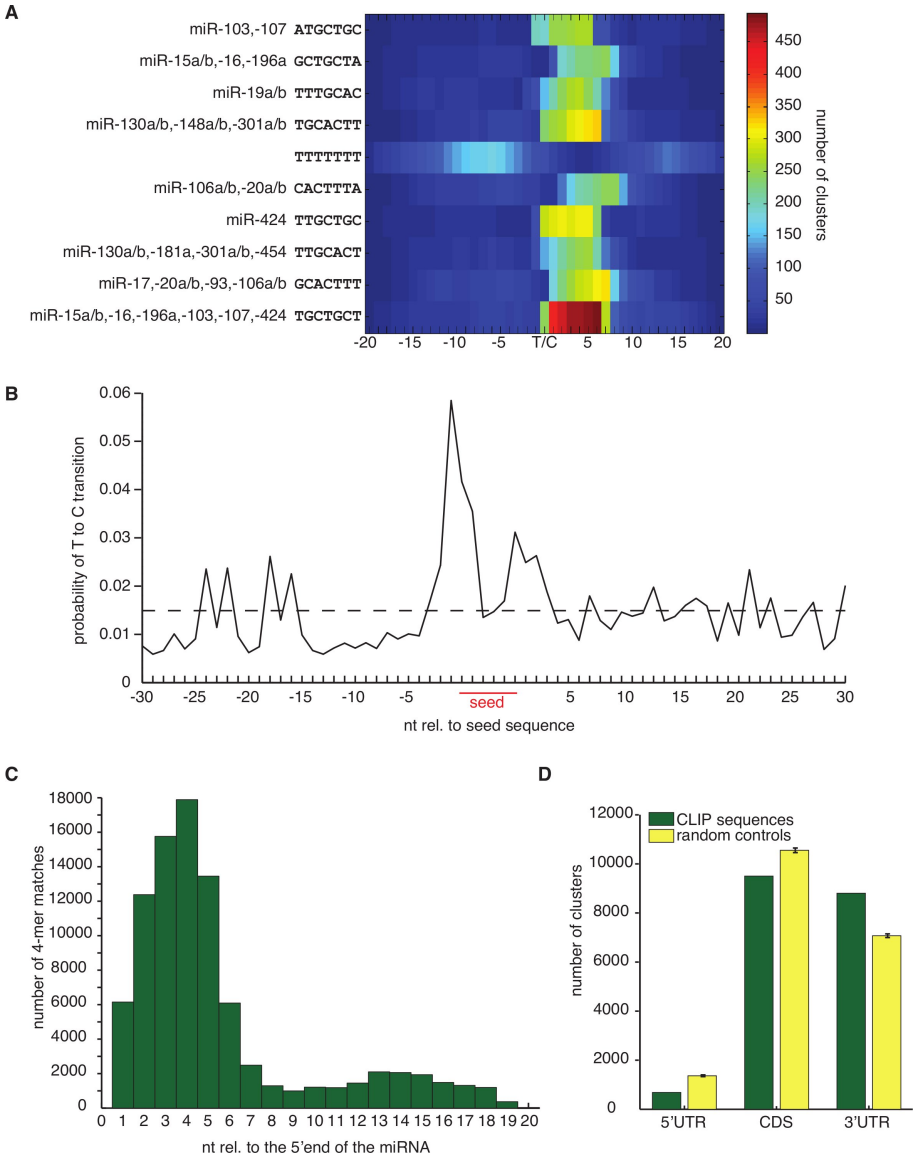


Figure 16: AGO PAR-CLIP identifies miRNA seed-complementary sequences in HEK293 cells. (A) Representation of the 10 most significantly enriched 7-mer sequences within PAR-CLIP CCRs. T/C indicates the predominant T to C transition within clusters of sequence reads. (B) T to C positional mutation frequency for clusters of sequence reads anchored at the 7-mer seed complementary sequence (pos. 2-8 of the miRNA) from all clusters containing seed-complementary sequences to any of the top 100 expressed miRNAs in HEK293 cells. The dashed line represents the average T to C mutation frequency within the clusters. (C) Identification of 4-nt base-pairing regions contributing to miRNA target recognition. CCRs with at least one 7-mer seed complementary region to one of the top 100 expressed miRNAs were selected. The number of 4-nt contiguous matches in the CCRs relative to the 5′ end of the matching miRNA was counted. (D) Analysis of the positional distribution of CCRs. The number of clusters annotated as derived from the 5′UTR, CDS or 3′UTR of target transcripts is shown (green bars). Yellow bars show the expected location distribution of the crosslinked regions if the AGO proteins bound without regional preference to the target transcript. See also Figure 58.

17,319 AGO CCRs, relative to random sequences with the same dinucleotide composition. The most significantly enriched 7-mers, except for a run of uridines, corresponded to the reverse complement of the seed region (position 2-8) of the most abundant HEK293 miRNAs, and they were most frequently positioned 1-2 nt downstream of the predominant crosslinking site within the CCRs (Figure 16A). This places the crosslinking site near the centre of the AGO-miRNA-target-RNA ternary complex, where the target RNA is proximal to the Piwi/RNase H domain of the AGO protein [227]. The polyuridine motif lies within the region of target RNA that may be able to basepair with the 3′ half of miRNA loaded into AGO proteins [227, 228]. Therefore, these stretches of uridine may contribute directly to miRNA-target RNA hybridization or, as has been suggested previously, they may represent an independent determinant of miRNA targeting specificity [83, 95].

To further examine the positional dependence of target RNA crosslinking, we aligned the CCRs containing 7-mer seed complements to the 100 most abundant miRNAs and plotted the position-dependent frequency of finding a crosslinked position (Figure 16B). This identified two additional crosslinking regions, which correspond to the unpaired 5′ and 3′ ends of the target RNA exiting from the AGO ternary complex, indicating that the window size of 41 nt centered on the predominant crosslink position always included the miRNA-complementary sites.

We then computed the number of occurrences of miRNA-complementary sequences of various lengths in the CCRs and calculated their enrichment (Table S6). The most significant enrichment was generally obtained with 8-mers that were complementary to miRNA seed regions (pos. 1-8). Inspection of the region between 3 nt upstream and 9 nt downstream of the predominant crosslinking site reveals that approximately 50% of the CCRs contain 6-mers corresponding to one of the top 100 expressed miRNAs (Figure 57), with a 1.5-fold enrichment over random 6-mers. Given that 6-mers still showed some degree of excess conservation in comparative genomics studies [70, 135] (Table S6) and that our analysis was focused on a narrow window directly downstream of the crosslinking site, our results suggest that the majority of the CCRs represent bona fide miRNA binding sites. Furthermore, the number of miRNA seed complements for all known miRNAs correlated well with the expression levels of miRNAs found in HEK293 cells, and less well with miRNA profiles of other tissue samples (Figure 58B). The nucleotide composition of CCRs that contained at least one 7-mer seed complementary to one of the top 100 expressed miRNA showed a slightly elevated U-content (approx. 30% U) compared to those CCRs not containing seed matches (Figure 58C), which was expected from previous bioinformatic analyses of functional miRNA-binding sites.

4.2.9 *Non-canonical and 3′end pairing of miRNAs to their mRNA targets is limited*

Structural and biochemical studies of the ternary complex of T. thermophilus Ago, guide and target indicated that small bulges and mismatches could be accommodated in the seed pairing region within the target RNA strand [227]. We therefore searched for putative target RNA binding sites that did not conform to the model of perfect miRNA seed pairing, but rather contained a discontinuous segment

of sequence complementarity to either target or miRNA with a minimum of 6 base pairs. We only considered pairing patterns if they were significantly enriched in CCRs compared to dinucleotide randomized sequences, and if the CCRs containing them did not at the same time contain perfectly pairing seed-type sites. We identified 891 CCRs with mismatches and 256 with bulges in the seed region (Table S7). Mismatches occurred most frequently across from position 5 of the miRNA as G-U or U-G wobbles, U-U mismatches and A-G mismatches (A residing in the miRNA). Therefore, it appears that only a small fraction of the miRNA target sites that we isolated (less than 6.6%), contained bulges or loops in the seed region.

To assess the role of auxiliary base pairing outside of the seed region, we selected CCRs that contained a 7-mer seed match to one of the 100 most abundant miRNAs. Supporting earlier computational results [83], we also detected a weak signal for contiguous 4-nt long matches to positions 13-15 of the miRNA (Figure 16C).

4.2.10 *miRNA binding sites in CDS and 3'UTR destabilize target mRNAs to different degrees*

The majority (84%) of AGO CCRs originated in exonic regions, with only 14% from intronic, and 2% from undefined regions. Of the exonic CCRs, 4% corresponded to 5'UTRs, 50% to CDS, and 46% to 3'UTRs (Figure 16D).

Evidence of widespread binding of miRNAs to the CDS was reported before [52, 135]. However, miRNAs are believed to predominantly act on 3'UTRs [13], with relatively few reports providing experimental evidence for miRNA-binding to individual 5'UTRs or CDS [52, 67, 145, 165, 208]. To obtain evidence that AGO CCRs indeed contain functional miRNA-binding sites, we blocked 25 of the most abundant miRNAs in HEK293 cells (Figure 15C) by transfection of a cocktail of 2'-O-methyl-modified antisense oligoribonucleotides and monitored the changes in mRNA stability by microarrays (Figure 17A). Consistent with previous studies of individual miRNAs [83], the magnitude of the destabilization effects of transcripts containing at least one CCR depended on the length of the seed-complementary region and dropped from 9-mer to 8-mer to 7-mer to 6-mer matches (Figure 17B). We did not find evidence for significant destabilization of transcripts that only contained imperfectly paired seed regions.

Next, we examined whether the change in stability of CCR-containing transcripts correlated with the number of binding sites. We found that multiple sites were more destabilizing compared to single sites (Figure 17C), and that multiple binding sites may also reside within a single 41-nt CCR (Figure 58). Both of these findings are in agreement with previous observations [83]. Then we analyzed the impact on stability for transcripts with CCRs exclusively present either in the CDS or the 3'UTR; there were not enough transcripts to assess the impact of CCRs derived from the 5'UTR. CDS-localized sites only marginally reduced mRNA stability (Figure 17D), independent of the extent of seed pairing. To gain more insights into miRNA binding in the CDS, we examined the codon adaptation index (CAI) [193] around crosslinked seed matches, and found that the sequence environment of crosslinked seed matches differed from that of non-crosslinked seed matches in the CAI. The bias in codon usage extended for at least 70 codons up-

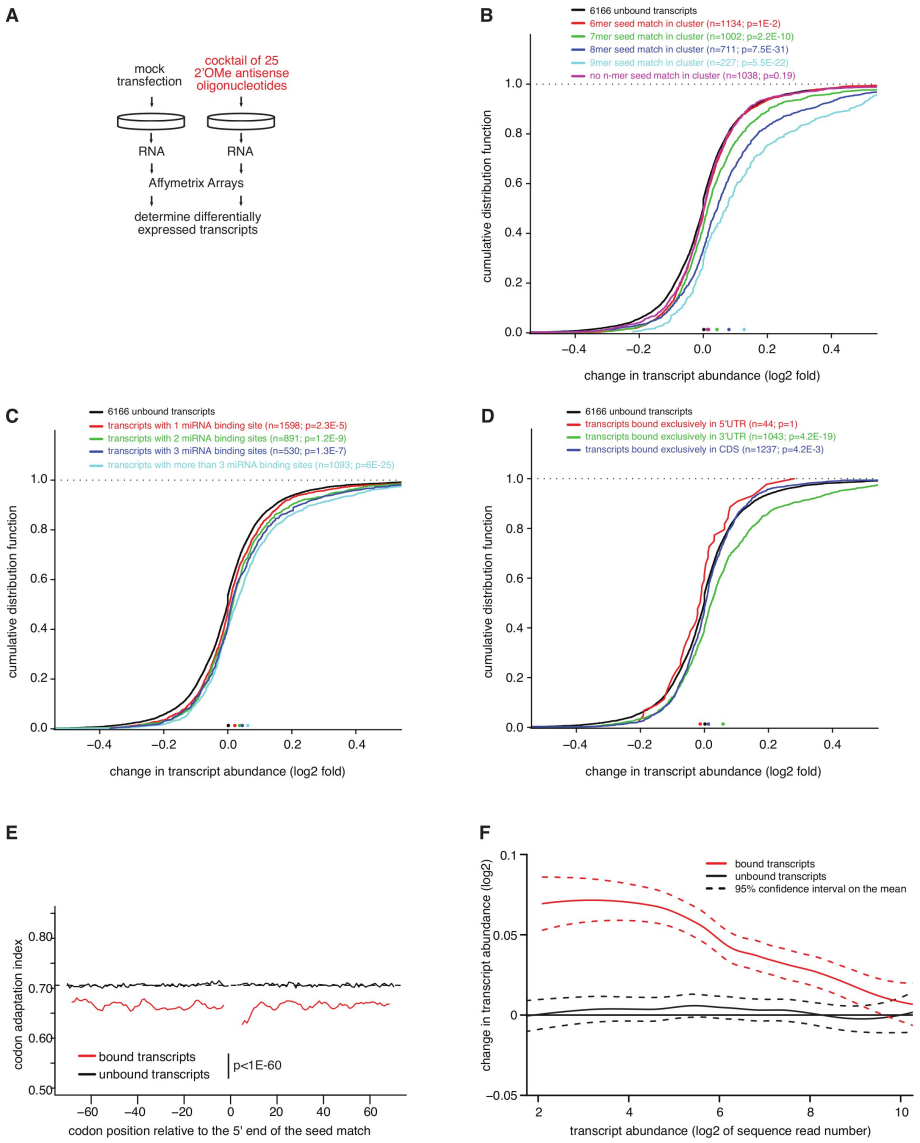


Figure 17: Relationship between various features of miRNA/target RNA interactions and mRNA stability (A) FLAG/HA-AGO2-tagged HEK293 cells were transfected with a cocktail of 25 2′-O-methyl modified antisense oligoribonucleotides, inhibiting miRNAs marked in red in Figure 15C, or mock transfected, followed by microarray analysis of the change of mRNA expression levels. (B) Transcripts containing CCRs were categorized according to the presence of n-mer seed complementary matches and the distributions of stability changes upon miRNA inhibition are shown for these categories. (C) Transcripts were categorized according to the number of CCRs they contained. (D) Transcripts were categorized according to the positional distribution of CCRs. Only transcripts containing CCRs exclusively in the indicated region are used. (E) Codon adaptation index (CAI) for transcripts containing 7-mer seed complementary regions (pos. 2-8) in the CDS for the miR-15, miR-19, miR-20, and let-7 miRNA families. (F) LOESS regression of total transcript abundance in HEK 293 cells (log2 of sequence counts determined by digital gene expression (DGE)) against fold change of transcript abundance (log2) determined by microarrays after transfection of the miRNA antagonist cocktail versus mock transfection of AGO-bound and unbound transcripts. See also Figure 59.

as well as downstream of the crosslinked seed matches (Figure 17E), which also correlates well with the marked increase in the A/U content around the binding sites that would lead to a codon usage bias. It was recently reported that miRNA regulation in the CDS was enhanced by inserting rare codons upstream of the miRNA-binding site, presumably due to increased lifetime of miRNA-target-RNA interactions as ribosomes are stalled [86]. These observations suggest that transcripts with reduced translational efficiency form at least transient miRNP complexes amenable to UV crosslinking.

The abundance of mRNAs expressed in HEK293 cells varied over 5 orders of magnitude as shown by DGE profiling. When we related the expression level of CCR-containing transcripts with the magnitude of transcript stabilization after miRNA inhibition, we found that miRNAs preferentially act on transcripts with low and medium expression levels (Figure 17F). Highly expressed mRNAs appear to avoid miRNA regulation [204], at least for those miRNAs expressed in HEK293 cells. However, we cannot fully rule out that the weaker response of highly abundant targets may be due to lower affinity and reduced occupancy of miRNA binding sites in highly abundant transcripts.

Earlier studies defining miRNA target regulation were carried out by transfection of miRNAs into cellular systems originally devoid of these miRNAs [8, 138, 191]. We transfected miRNA duplexes corresponding to the deeply conserved miR-7 and miR-124 into FLAG/HA-AGO2 cells, performed PAR-CLIP (Figure 59), and also recorded the effect on mRNA stability upon miR-7 and miR-124 transfection by microarray analysis. Transcripts containing miR-7- or miR-124-specific CCRs were destabilized, especially when CCRs were located in the 3'UTR (Figure 59).

4.2.11 *Context-dependence of miRNA binding*

Not every seed-complementary sequence in the HEK293 transcriptome yielded a CCR, thereby providing an opportunity to identify sequence context features specifically contributing to miRNA target binding and crosslinking. For seed-complementary sites that were crosslinked and those that were not crosslinked, we computed the evolutionary selection pressure by the ElMMo method [70], the mRNA stability scores by TargetScan context score [83], and sequence composition and structure measures for the regions around the miRNA seed complementary sites. The feature that distinguished most crosslinked from non-crosslinked seed matches was a 25% lower free energy required to resolve local secondary structure involving the miRNA-binding region (Figure 59), associated with a 6% increase in the A/U content within 100 nt around the seed-pairing site. These differences were similar for sites located in the CDS and 3'UTRs. Compared to non-crosslinked sites, crosslinked sites are under stronger evolutionary selection (ElMMo) and in sequence contexts facilitating miRNA-dependent mRNA degradation (TargetScan context score).

The location of AGO CCRs within transcript regions was non-random and 7-mer or 8-mer sites within the 3'UTR were preferentially located near the stop codon or the polyA tail in transcripts with relatively long 3'UTRs (more than 3 kb) (Figure 59). The location of CCRs in the CDS was biased towards the stop codon for the transfected miR-7 and 124, but not for the endogenous miRNAs (Figure 59).

Finally, we wanted to examine how miRNA targets defined by PAR-CLIP compared in regulation of target mRNA stability to those predicted by ElMMo [70], TargetScan context score [83], TargetScan Pct [69] and PicTar [128]. In each case, we selected the same number of highest-scoring sites containing a 7-mer seed-complement to the top 5 expressed miRNAs (let-7a, miR-103, miR-15a, miR-19a and miR-20a). The analysis was limited to 3'UTR sites due to restriction by the prediction methods. The effect on mRNA stability, as assessed by miRNA antisense inhibition, was overall equivalent for transcripts harboring CCRs compared to transcripts predicted by ElMMo, TargetScan context score, TargetScan Pct and PicTar (Figure 59).

4.3 DISCUSSION

Maturation, localization, decay and translational regulation of mRNAs involve formation of complexes of RBPs and RNPs with their RNA targets [147, 156]. Several hundred RBPs are encoded in the human genome, many of them containing combinations of RNA-binding domains which are drawn from a relatively small repertoire, resulting in diverse structural arrangements and different specificities of target RNA recognition [143]. Furthermore hundreds of miRNAs function together with AGO and TNRC6 proteins to destabilize target mRNAs and/or repress their translation [13]. Collectively, these factors and their presumably combinatorial action constitute the code for post-transcriptional gene regulation. Here we describe an approach to directly identify transcriptome-wide mRNA-binding sites of regulatory RBPs and RNPs in live cells.

4.3.1 *PAR-CLIP allows high-resolution mapping of RBP and miRNA target sites*

We showed that application of photoactivatable nucleoside analogs to live cells facilitates RNA-protein crosslinking and transcriptome-wide identification of RBP and RNP binding sites. We concentrated on 4SU after it became apparent that the crosslinking sites in isolated RNAs were revealed upon sequencing by a prominent transition from T to C in the cDNA prepared from the isolated RNA segments. Compared to regular UV 254 nm crosslinking in the absence of photoactivatable nucleosides, our method has two distinct advantages. We obtain higher yields of crosslinked RNAs using similar radiation intensities, and more importantly, we can identify crosslinked regions by mutational analysis. Studies using conventional UV 254 nm CLIP have not reported the incidence of deletions and substitutions [38, 136, 214, 241], except for recent work by Granneman et al. [80] on the U3 snoRNA that showed an increase of deletions at the RBP binding site. Our own analysis indicates that mutations in sequence reads derived from UV 254 nm CLIP were at least one order of magnitude less frequent than T to C transitions observed in PAR-CLIP (Figure 55).

From an experimental perspective, it is important to note that crosslinked RNA segments, irrespective of the methods of isolation, are always contaminated with non-crosslinked RNAs, as shown by consistent identification of rRNAs, tRNAs, and miRNAs (Table S2). Compared to crosslinked RNA fragments, these unmodified RNA molecules are more readily reverse transcribed, which underscores the need for sep-

aration of crosslinked signal from non-crosslinked noise. We now provide a method that accomplishes this critical task.

4.3.2 *Context dependence of 4SU crosslink sites*

It is conceivable that binding sites located in peculiar sequence environments, e.g. those completely devoid of U, may exist and cannot be captured using 4SU-based crosslinking. However, such sites are extremely rare. Only about 0.4% of 32-nt long sequence segments, representative of the length of our Solexa sequence reads, are U-less, corresponding to an occurrence of one such segment in every 8 kb of a transcript.

Nonetheless, to provide a means to resolve such unlikely situations, we explored the use of other photoactivatable nucleosides, such as 6SG to identify IGF2BP1 binding sites. We found a good correlation between the sequence reads obtained from a given gene with 4SU and 6SG (Pearson correlation coefficient 0.65, Table S1). Moreover, the sequence read clusters, representing individual binding sites, overlapped strongly: 59% out of the 47,050 6SG clusters were also identified with 4SU, despite of the fact that the environment of IGF2BP1 binding sites was strongly depleted for guanosine. Interestingly, the sequence reads obtained after 6SG crosslinking were enriched for G to A transitions, pointing to a structural change in 6SG analogous to the situation in PAR-CLIP with 4SU. Because 6SG appears to have lower crosslinking efficiency compared to 4SU, we recommend to first use 4SU and then resort to 6SG when the data indicates that the sites of interest are located in sequence contexts devoid of uridines. It is important to point out that neither of these photoactivatable nucleotides appears to be toxic under our recommended conditions.

4.3.3 *miRNA target identification*

When applying PAR-CLIP to isolate miRNA-binding sites, we were surprised to find nearly 50% of the binding sites located in the CDS. However, miRNA inhibition experiments showed that miRNA binding at these sites only caused small, yet significant mRNA destabilization. In spite of the difference in their efficiency of triggering mRNA degradation, CDS and 3'UTR sites appear to have similar sequence and structure features. The sequence bias around CDS sites is associated with an increased incidence of rare codon usage, which could in principle reduce translational rate, thereby providing an opportunity for transient miRNP binding and regulation. Similar observations were made previously using artificially designed reporter systems [86].

The use of the knowledge of the crosslinking site allowed us to narrowly define the miRNA-binding regions for matching the site with the most likely miRNA endogenously co-expressed with its targets, and to assess non-canonical miRNA-binding modes. We were able to explain the majority of PAR-CLIP binding sites by conventional miRNA-mRNA seed-pairing interactions [83], yet found that about 6% of miRNA target sites might best be explained by accepting bulges or mismatches in the seed pairing region, similar to the interaction between let-7 and its target lin-41 [221] and those recently observed in biochemical and structural studies of T. thermophilus Ago protein [227, 228].

4.3.4 *The mRNA ribonucleoprotein (mRNP) code and its impact on gene regulation*

We were able to identify all of the crosslinkable RNA-binding sites present in about 9,000 of the top-expressed mRNA in HEK293 cells representing approximately 95% of the total mRNA molecules of a cell. One of the surprising outcomes of our study was that each of the examined RBPs or miRNPs bound and presumably controlled between 5 and 30% of the more than 20,000 transcripts detectable in HEK293 cells. These results demonstrate that a transcript will generally be bound and regulated by multiple RBPs, the combination of which will determine the final gene-specific regulatory outcome. Exhaustive high-resolution mapping of RBP- and RNP-target-RNA interactions is critical, because it may lead to the discovery of specific combination of sites (or modules) that may control distinct cellular processes and pathways. To gain further insights into the dynamics of mRNPs it will be important to also map the sites of RNA-binding factors, such as helicases, nucleases or polymerases, where the specificity determinants are poorly understood. The precise identification of RNA interaction sites will be extremely useful for interrogating the rapidly emerging data on genetic variation between individuals and whether some of these variations possibly contribute to complex genetic diseases by affecting post-transcriptional gene regulation.

4.4 METHODS

4.4.1 *PAR-CLIP*

Human embryonic kidney (HEK) 293 cells stably expressing FLAG/HA-tagged IGF2BP1-3, QKI, PUM2, AGO1-4, and TNRC6A-C [130] were grown overnight in medium supplemented with 100 μM 4SU. Living cells were irradiated with 365 nm UV light. Cells were harvested and lysed in NP40 lysis buffer. The cleared cell lysates were treated with RNase T1. FLAG/HA-tagged proteins were immunoprecipitated with anti-FLAG antibodies bound to Protein G Dynabeads. RNase T1 was added to the immunoprecipitate. Beads were washed and resuspended in dephosphorylation buffer. Calf intestinal alkaline phosphatase was added to dephosphorylate the RNA. Beads were washed and incubated with polynucleotide kinase and radioactive ATP to label the crosslinked RNA. The protein-RNA complexes were separated by SDS-PAGE and electroeluted. The electroeluate was proteinase K digested. The RNA was recovered by acidic phenol/chloroform extraction and ethanol precipitation. The recovered RNA was turned into a cDNA library as described [89] and Solexa sequenced. The extracted sequence reads were mapped to the human genome (hg18), human mRNAs and miRNA precursor regions. For a more detailed description of the methods, see the Supplementary Material.

4.4.2 *Oligonucleotide transfection and mRNA array analysis*

siRNA, miRNA and 2'-O-methyl oligonucleotide transfections of HEK293 T-REx Flp-In cells were performed in 6-well format using Lipofectamine RNAiMAX (Invitrogen) as described by the manufacturer. Total RNA of transfected cells was extracted using TRIZOL following the in-

structions of the manufacturer. The RNA was further purified using the RNeasy purification kit (Qiagen). 2 µg of purified total RNA was used in the One-Cycle Eukaryotic Target Labeling Assay (Affymetrix) according to manufacturer's protocol. Biotinylated cRNA targets were cleaned up, fragmented, and hybridized to Human Genome U133 Plus 2.0 Array (Affymetrix). For details of the analysis, see Bioinformatics section in the Supplementary Material.

4.4.3 *Generation of Digital Gene Expression (DGEX) libraries*

1 µg each of total RNA from HEK293 cells inducibly expressing tagged IGF2BP1 before and after induction was converted into cDNA libraries for expression profiling by sequencing using the DpnII DGE kit (Illumina) according to instructions of the manufacturer. For details of the analysis, see Bioinformatics section in the Supplementary Material.

4.5 ACKNOWLEDGMENTS

We thank V. Hovestadt for his help with the analysis of the crosslinking positions within miRNAs. We are grateful to W. Zhang and C. Zhao (Genomics Resource Center) for mRNA array analysis and Solexa sequencing. We thank Millipore for the antibodies. We thank members of the Tuschl laboratory for comments on the manuscript. M.H. is supported by the Deutscher Akademischer Austauschdienst (DAAD). This work was supported by the Swiss National Fund Grant #3100A0-114001 to M.Z.; T.T. is an HHMI investigator, and work in his laboratory was supported by NIH grants GM073047 and MH08442 and the Starr Foundation.

T.T. is a cofounder and scientific advisor to Alnylam Pharmaceuticals and an advisor to Regulus Therapeutics.

5

MIRZ: AN INTEGRATED MICRORNA EXPRESSION ATLAS AND TARGET PREDICTION RESOURCE

The work presented in this chapter was originally published in Nucleic Acids Research [94]

5.1 INTRODUCTION

Studies in both native expression [130] as well as transfection-induced miRNA overexpression situations [138, 140] indicate that within a given tissue, the miRNAs that are most strongly expressed have the largest impact on mRNA targets. For this reason, deciphering the miRNA-dependent post-transcriptional regulatory layer in a given tissue or cell type needs to start from the miRNA expression profile of that tissue or cell type. Conversely, it is very common that one identifies differences in miRNA expression between cells at various stages of differentiation or between normal and malignant cells, and the natural question is what mRNAs are most likely to be affected by the change in miRNA expression. To address these types of questions, we developed MirZ (`www.mirz.unibas.ch`), a web service that integrates two resources that we developed in the context of previous research projects: the smiRNAdb miRNA expression atlas [129], and the ElMMo miRNA target prediction algorithm [70].

5.2 MATERIALS AND METHODS

5.2.1 *The smiRNAdb miRNA expression atlas*

smiRNAdb [129] is a web-accessible and widely used resource of miRNA profiles determined by sequencing from hundreds of *Homo sapiens*, *Mus musculus* and *Rattus norvegicus* samples. The miRNA expression profiling approach used by smiRNAdb is small RNA sequencing by classical cloning and sequencing of size-separated small RNAs, which was used to generate a large atlas of miRNA expression profiles [129]. This approach can be scaled up considerably through deep sequencing technologies [15, 89, 90]. Microarray-based expression profiling is also a popular approach, which has been used for instance to characterize the miRNA expression cancer samples [142]. In contrast to sequencing, microarray-based profiling does not allow identification of novel miRNAs.

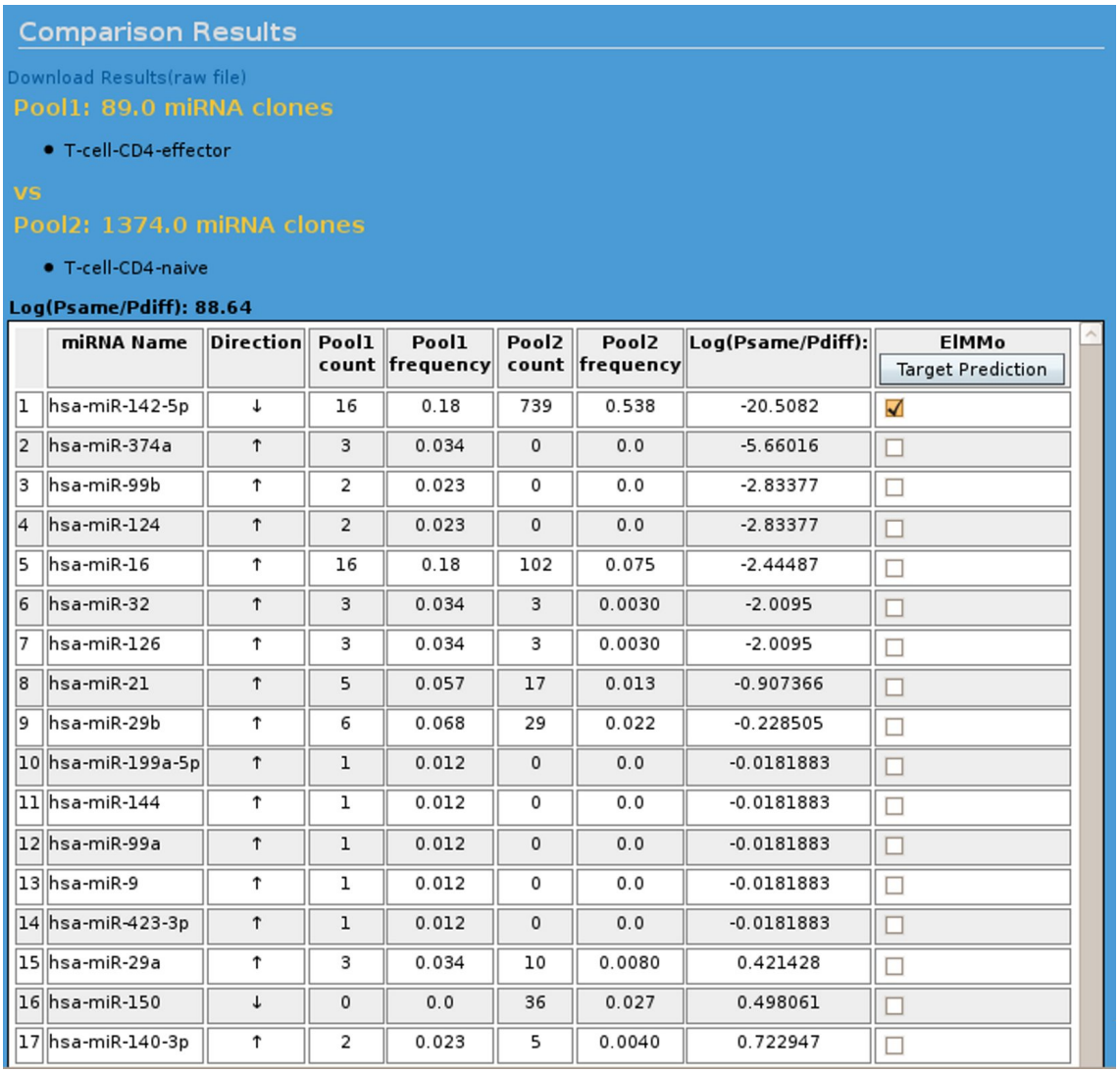
The web interface of smiRNAdb features an extended repertoire of on-line analyses such as visualization and hierarchical clustering of miRNA expression profiles, principal component analysis, comparison of miRNA expression between two (sets of) samples with the aim of identifying the miRNAs whose expression differs most between the samples. We used the Brenda tissue ontology [189, 11] as a guide in organizing the samples such that the user can readily identify related cell lineages or normal and pathological samples derived from a given tissue type. Our tissue hierarchy has four levels: the organ/system (e.g. hematopoietic system), subsystem (e.g. lymphoid lineage), cell type (e.g. B cell), further cell type classification (e.g. B lymphocyte). MiRNAs themselves can be analyzed independently, grouped by their 2-7 subsequence, or grouped in precursor clusters. Two miRNAs are

placed in the same precursor cluster if their loci are within 50 kilobases of each other in the genome, or if they share a mature form.

As an example, one may be interested in comparing miRNA expression between effector and naive human $CD4^+$ T-lymphocytes. SmiRNAdb features a "Sample comparison" tool which was specifically designed for the pairwise comparison of miRNA (sets of) samples. The user would select to compare the sample named "hsa_T-cell-CD4-effector" to the sample named "hsa_T-cell-CD4-naive". Because the naive $CD4^+$ T cell sample and the effector $CD4^+$ T cell sample differ widely in the total number of sequenced miRNAs (1374 vs 89), the precision of the miRNA frequency estimates in the two samples will also be very different. This situation is common in sequencing-based datasets making the identification of miRNAs whose expression is *significantly* different a non-trivial problem. At the heart of the tools offered by smiRNAdb however, is a Bayesian model for computing the posterior probability that the *frequency* of a miRNA in the total miRNA population differs between two (sets of) samples. We compute this probability assuming a binomial sampling model and integrating over the unknown miRNA frequencies in the samples. This approach — described in details in Berninger et al. [16] — takes into account both the variability between sample sizes and the absolute miRNA counts.

Figure 18 shows the results of comparing the miRNA expression profiles of naive *vs* effector $CD4^+$ cells. The names and sizes of the samples being compared are shown at the top of the page, followed by the log-likelihood ratio $\log(P_{same}/P_{diff})$ of two models, one that assumes that the frequencies of miRNAs are the same and one that assumes that they can be different between the samples. The log-likelihood ratio takes positive values when the miRNA frequencies are similar and negative values when they are different. In this case, the log-likelihood ratio is positive, indicating that overall, the frequencies of miRNAs in these samples are more likely to have been the same. The list of miRNAs ranked from most dissimilar to most similar expression follows. Each row contains the name of a miRNA, the direction of regulation (up or down), the cloning counts and frequencies in both samples, and provides a direct link to the predicted targets of the miRNA. The model indicates that with a 18% vs 54% cloning frequency, and despite the small size of the effector $CD4^+$ T cell sample, miR-142-5p is very likely to be down-regulated in effector cells. Again, this can be inferred from the negative value of $\log(P_{same}/P_{diff})$ for miR-142-5p. From this page, the user can select one or several miRNAs that came out differentially expressed and can browse the list of predicted targets (figure 19). In the case of miR-142-5p, the top 10 predicted targets include four transcription factors (AFF4, ONECUT2, ZFPM2 and ZNF148), and a kinase (PRPF4B) involved in pre-RNA splicing. These genes could provide a starting point for experimental studies on the function of miR-142-5p in T lymphocytes.

Since the original release of smiRNAdb, we have implemented an additional tool for performing principal component analysis on the miRNA expression profiles, we added more possibilities for the user to download miRNA profile data for further processing, and we started to incorporate other publicly available small RNA sequencing data sets from *Danio rerio*, *Drosophila melanogaster* and *Caenorhabditis elegans*. We reimplemented the software that was originally written in Perl CGI to use Java Server Faces technology and Apache / Tomcat. The com-



Comparison Results

Download Results(raw file)

Pool1: 89.0 miRNA clones

- T-cell-CD4-effector

vs

Pool2: 1374.0 miRNA clones

- T-cell-CD4-naive

Log(Psame/Pdiff): 88.64

	miRNA Name	Direction	Pool1 count	Pool1 frequency	Pool2 count	Pool2 frequency	Log(Psame/Pdiff):	ElMMo Target Prediction
1	hsa-miR-142-5p	↓	16	0.18	739	0.538	-20.5082	☑
2	hsa-miR-374a	↑	3	0.034	0	0.0	-5.66016	☐
3	hsa-miR-99b	↑	2	0.023	0	0.0	-2.83377	☐
4	hsa-miR-124	↑	2	0.023	0	0.0	-2.83377	☐
5	hsa-miR-16	↑	16	0.18	102	0.075	-2.44487	☐
6	hsa-miR-32	↑	3	0.034	3	0.0030	-2.0095	☐
7	hsa-miR-126	↑	3	0.034	3	0.0030	-2.0095	☐
8	hsa-miR-21	↑	5	0.057	17	0.013	-0.907366	☐
9	hsa-miR-29b	↑	6	0.068	29	0.022	-0.228505	☐
10	hsa-miR-199a-5p	↑	1	0.012	0	0.0	-0.0181883	☐
11	hsa-miR-144	↑	1	0.012	0	0.0	-0.0181883	☐
12	hsa-miR-99a	↑	1	0.012	0	0.0	-0.0181883	☐
13	hsa-miR-9	↑	1	0.012	0	0.0	-0.0181883	☐
14	hsa-miR-423-3p	↑	1	0.012	0	0.0	-0.0181883	☐
15	hsa-miR-29a	↑	3	0.034	10	0.0080	0.421428	☐
16	hsa-miR-150	↓	0	0.0	36	0.027	0.498061	☐
17	hsa-miR-140-3p	↑	2	0.023	5	0.0040	0.722947	☐

Figure 18: Screenshot of the web page showing the result from comparing miRNA expression of human $CD4^+$ effector T cells with the $CD4^+$ naive T cells. Details are provided in the text.

putations are now performed on a computing cluster, with job distribution managed by the Sun Grid Engine queuing system. Finally, we enhanced the result screens of our on-line analysis tools with hyperlinks which directly take the user to the miRNA target predictions within the context of the smiRNAdb query, *i.e.* preserving the selected organism, miRNAs, and tissue (if available). Please refer to the web connectivity map in the supplementary material for an overview of the new links between smiRNAdb and ElMMo, as well as of the external ressources that we use in performing various analyses.

5.2.2 *The ElMMo miRNA target prediction algorithm based on comparative genomic analysis*

To be able to address the question of what mRNA is most likely affected by the change in expression of a miRNA, we coupled smiRNAdb to a PHP-based web interface to the ElMMo miRNA target predictions [70].

Returning to the example of the hsa-miR-142-5p miRNA which was highlighted in section 5.2.1, the web interface allows aside from browsing the predicted targets, a number of other queries. For instance, given an organism (*Homo sapiens* in this example), the user can choose to scan for predicted miRNA target sites not only the default set of transcripts, which is all known RefSeq [172] mRNAs in the chosen organism, but also subsets of transcripts. The SymAtlas project [205] of the Genomics Institute of the Novartis Research Foundation (GNF) generated microarray-based *mRNA* expression profiles for a wide range of tissues. These profiles are incorporated in MirZ, giving the user the possibility to restrict target prediction to mRNAs that are expressed in a given cell type. The web interface further allows to scan an arbitrary number of mRNAs for up to 20 miRNAs simultaneously. Alternatively, the user can limit the number of mRNAs to scan to 20 mRNAs and then retrieve predicted target sites in these mRNAs for an arbitrary number of miRNAs.

MiRNAs exert their effector function through ribonucleoprotein complexes (miRNP) that contain, aside from the guiding miRNA a member of the Argonaute family of proteins. The determinants of productive miRNA-target site interactions are not entirely known, but a large body of work [127, 134, 47, 176, 135, 24] established that perfect complementarity of the 7–8 nucleotides from the 5′ end of the miRNA — the so-called miRNA "seed" — is critical for target recognition. Although miRNA target sites that do not satisfy this constraint have been described, at the genome-wide level the accuracy of predicting such sites is low [135, 70]. Other than perfect seed complementarity, the location of the putative target site within the 3′ UTR [70, 83, 146], structural accessibility [141, 117, 207], the nucleotide composition in its vicinity [83, 161] and the complementarity of specific positions in the miRNA 3′ end to the target site [83] have all been reported to improve the accuracy of miRNA target prediction, yet the relative importance of these features was unknown until we performed the work described in Chapter 3. Therefore, miRNA-mRNA association were obtained computationally using the ElMMo miRNA target prediction method developed in the Zavolan lab, which is based on a Bayesian model that only uses comparative genomics information [70]. Ongoing software development efforts in the Zavolan lab aim at generalizing mirz to ClipZ,

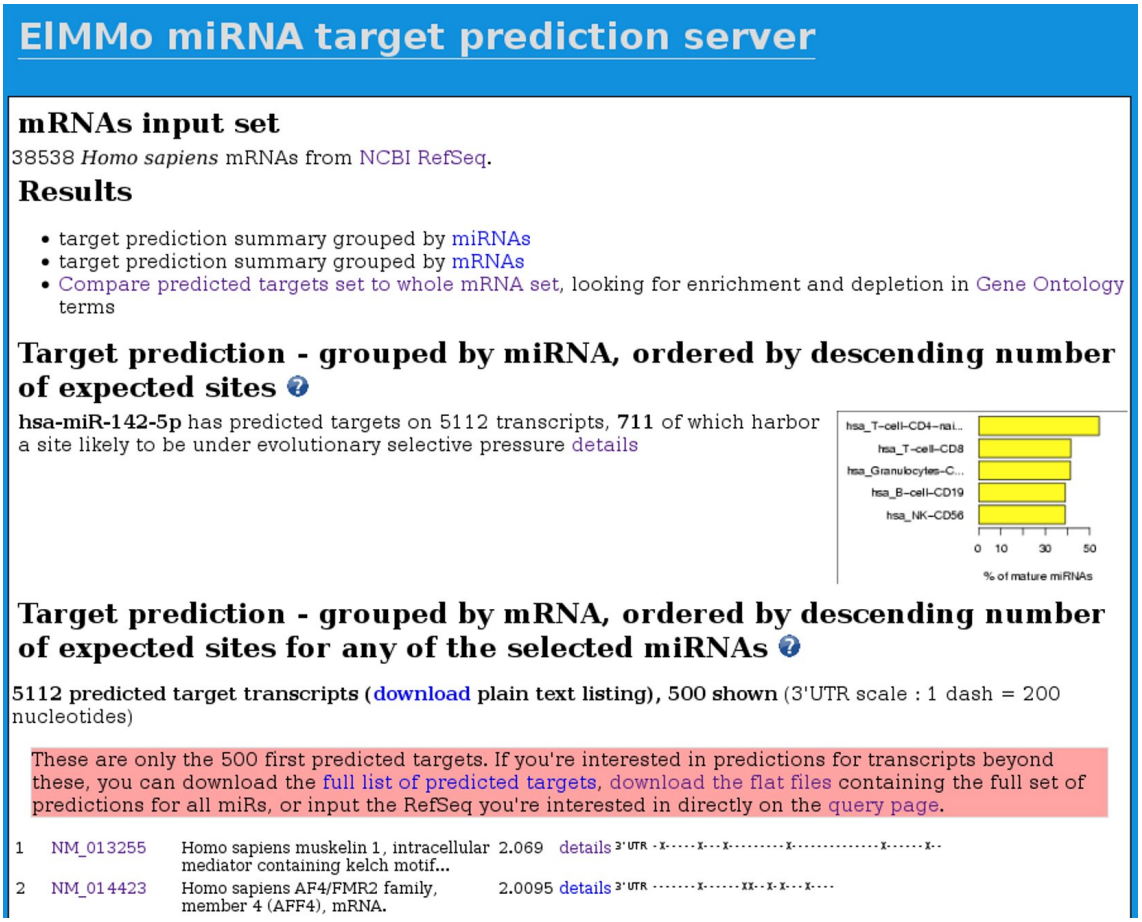


Figure 19: Screenshot of the web page showing the ElMMo miRNA target predictions for miR-142-5p in all *Homo sapiens* RefSeq mRNAs. The target predictions results are organized in two sections. The first section — located on the upper part of the web page — is miRNA-centric and features miRNA target predictions statistics as well as a figure showing the smiRNAdb tissues where the miRNA is mostly expressed. The second, mRNA-centric section is located on the lower part of the web page and provides a ranked list of mRNA predicted to be targeted by miR-142-5p.

which is a service that enables the exploration of miRNA - mRNA, RNA binding protein - mRNA and miRNA profiles in an integrated environment.

Going back to our example, figure 19 shows the ElMMo predictions for miR-142-5p in *Homo sapiens*. This result screen is organized in two sections: (1) a miRNA-centric summary featuring per-miRNA target prediction statistics and a figure showing the smiRNAdb tissues where the selected miRNAs are mostly expressed, and (2) a mRNA-centric summary that ranks all mRNAs predicted to be targeted by the selected miRNAs. In this later section, mRNAs are ordered by decreasing expected number of miRNA target sites under selective pressure, defined as the sum of all target site posterior probabilities for the selected miRNAs. The location of the putative target sites in the 3'UTR is also indicated.

From the result screen, the user has the possibility to zoom onto a specific transcript to visualize the multiple genome alignments in the

regions of the predicted target sites, and to find additional information about the targeted mRNAs from the Genbank database of the National Center for Biomedical Information (NCBI). Our web service also offers the possibility to run a Gene Ontology (GO) analysis searching for GO terms that are significantly over- or under-represented in the predicted miRNA targets through a modified version of the GeneMerge software [32]. For instance, in the case of miR-142-5p, the most significantly enriched Biological Process GO term is "regulation of transcription, DNA-dependent" (hypergeometric p-value $< 10^{-10}$, after Bonferroni multiple testing correction), followed by two "muscarinic acetylcholine receptor"-associated GO terms ($p < 10^{-10}$). The muscarinic acetylcholine receptor has been shown to be involved in autocrine control of cell proliferation, including the proliferation of immune cells [55]. This type of analyses could thus provide experimental scientists with clues to the function of miR-142-5p in the naive $CD4^+$ T cells.

The current release of ElMMo features miRNA target predictions for *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Caenorhabditis elegans* and *Drosophila melanogaster*. Of these, the *Mus musculus* and *Rattus norvegicus* predictions were not present in our initial publication [70]. Furthermore, for the remaining organisms, the current predictions are based on the genome sequences of a larger set of species, because more fully-sequenced genomes became available since 2007. We further based our predictions on the most recent mRNA sequences and 3'UTR annotations provided by the RefSeq database [172]. Concerning the microarray profiles that the user can use to guide miRNA target discovery in specific tissues and aside from the *Homo sapiens* profiles that were used in our original ElMMo release [70], we incorporated similar mRNA expression profiles for *Mus musculus* and *Rattus norvegicus*. Finally, the ElMMo web interface now informs the user about the smiRNAdb samples in which the selected miRNAs are most strongly expressed.

5.2.3 *Experimental data*

The miRNA sequences that were used for miRNA sample annotation and for miRNA target prediction were obtained from the miRBase release 12.0 [82]. For the miRNA profiles, MirZ includes a total of 297 samples: 173 for *Homo sapiens* [129], 88 for *Mus musculus* [129], 16 for *Rattus norvegicus* [129], 10 for *Drosophila melanogaster* [6], 9 for *Danio rerio* [35], and 1 for *Caenorhabditis elegans* [184].

For miRNA target predictions, we used the most recent genome assemblies available at the University of California Santa Cruz (UCSC) [111]: hg18 for *Homo sapiens*, mm9 for *Mus musculus*, rn4 for *Rattus norvegicus*, danRer5 for *Danio rerio*, ce6 for *Caenorhabditis elegans* and dm3 for *Drosophila melanogaster*. We further used the following UCSC genome assemblies in the pairwise genome alignments: panTro2, rheMac2, mm9, rn4, canFam2, monDom4, bosTau4 and galGal3 for *Homo sapiens*; panTro2, rheMac2, hg18, rn4, canFam2, monDom4, bosTau4 and galGal3 for *Mus musculus*; panTro2, rheMac2, hg18, mm9, canFam2, monDom4, bosTau3 and galGal3 for *Rattus norvegicus*; tetNig1, fr2 and oryLat2 for *Danio rerio*; caeJap1, caePb2, caeRem3, cb3 and priPac1 for *Caenorhabditis elegans*; dp4, droAna3, droEre2, droGri2, droMoj3, droPer1, droSec1, droSim1, droVir3, droWil1 and droYak2 for *Drosophila melanogaster*.

mRNAs for all organisms were downloaded from the RefSeq database on January 21st 2009.

The links between sequence entities in various databases was made by mapping them all to the Gene database of NCBI [172]. MiRNA expression profiles, microarray mRNA profiles and miRNA target predictions are stored as relational databases managed by a PostgreSQL server (www.postgresql.org).

5.3 CONCLUSION AND FUTURE DIRECTIONS

Using a concrete example comparing effector to naive $CD4^+$ T-cells, we showed how MirZ can help isolating miRNAs that may be involved in a given biological function, and then provide clues into which molecular pathways may be controlled by these miRNAs to achieve their biological function. The integration of miRNA expression profiles with genome-wide miRNA target prediction combined with the tools we implemented — a Bayesian model for sample comparison, multivariate exploratory statistics, GO-term enrichment analysis — makes MirZ a powerful tool for studying miRNA-based regulation.

Since its publication, the miRNA expression atlas has been a valuable resource to the research community, and with the more general availability of deep sequencing technologies, more miRNA expression data sets are expected to emerge. Being able to explore and compare these data sets in a unified framework is highly desirable, and we plan to further support such analyses by updating MirZ as new data sets become available. Particularly for *Drosophila melanogaster*, we currently only incorporate small-sized samples, and for *Caenorhabditis elegans* a whole-worm sample.

The target prediction methods also continue to evolve. In particular, Chapter 3 examined what additional determinants of miRNA targeting specificity can be used to predict functional miRNA binding sites with better accuracy. The predictions from the corresponding model could be incorporated in our server in the future.

ACKNOWLEDGEMENTS

We acknowledge the International Chicken Genome Sequencing Consortium [101] for the *Gallus gallus* genome, the Chimpanzee Genome Sequencing Consortium for the *Pan troglodytes* genome, the Baylor College of Medicine Human Genome Sequencing Center and the Rhesus Macaque Genome Sequencing Consortium for the *Macaca mulatta* genome, the Mouse Genome Sequencing Consortium for the *Mus musculus* genome [229], the Rat Genome Project at the Baylor College of Medicine Human Genome Sequencing Center for the *Rattus norvegicus* genome [96, 77], the Dog Genome Sequencing Project for the *Canis familiaris* genome [139], the Broad Institute for the *Monodelphis domestica* (opossum) genome, the Baylor College of Medicine Human Genome Sequencing Center for the *Bos taurus* and *Drosophila pseudoobscura* genomes [42], the Genoscope for the *Tetraodon nigroviridis* genome, the Morishita laboratory for the *Oryzias latipes* (Medaka) genome [112], WormBase for the *Caenorhabditis elegans* genome [18], the Agencourt Bioscience Corporation for the *Drosophila ananassae* and *Drosophila erecta* genomes, and the Drosophila 12 Genomes Consortium for all other *Drosophila* genomes [39]. The *Takifugu rubripes* genome was provided

freely by the Fugu Genome Consortium for use in this publication only. The *C. brenneri, C. briggsae, C. japonica, C. remanei* and *P. pacificus* genomes were produced by the Genome Sequencing Center at Washington University School of Medicine in St. Louis and can be obtained from ftp://genome.wustl.edu/pub/organism/Invertebrates/.

6

TOWARD A KINETIC MODEL OF MICRORNA-MEDIATED GENE SILENCING AT THE MRNA AND PROTEIN LEVEL

6.1 INTRODUCTION

The temporal aspects of miRNA regulation are critical in several domains of biology. A canonical example is the control of lateral hypodermal cell lineage by the let-7 miRNA during C. elegans development [178]. In wild-type, let-7 expression is initiated at the L3 stage, which results in the silencing of the lin-41 gene, which in turns stops cell division. In the absence of let-7, the cells don't differentiate. Instead, they keep deviding until they burst through the vulva. A bad timing in let-7 regulation may therefore be lethal to C. elegans.

A second example where the temporal dynamics of miRNA regulation are critical is the miRNA target identification problem, which was extensively discussed in chapter 3. The discussion of this chapter left one question unanswered. By analyzing measurements of changes in protein and mRNA levels following miRNA transfection, we could show that structural accessibility, the AU content of the target mRNA and a few other properties characterize potent miRNA binding sites. However, we were very surprised to find that the same properties systematically failed to characterize miRNA binding sites that lead to down-regulated protein levels. The cause of that inconsistency is quite an enigma, because changes in protein levels should be the ultimate read-out of miRNA regulation in the sense that any change at the mRNA level should at least propagate down to the cognate protein, if not be amplified by concordant diminution of the translation rate [99]. It has been proposed recently that miRNA mostly up-regulate mRNA decay, leaving translation unchanged [88]. But even under this hypothesis, one should observe that miRNA binding sites leading to down-regulated protein levels have properties similar to miRNA binding-sites leading to mRNA degradation and miRNA binding-sites under evolutionary selective pressure.

One could argue that the measurements from proteomics experiments are not accurate enough to measure miRNA action. But the analysis suggests this is not the case, as there is statistical evidence that proteins encoded by mRNAs carrying miRNA binding sites are down-regulated following the over-expression of the cognate miRNA (see Selbach et al. [191], Baek et al. [8] and Fig. 40). Another possible explanation could be the sample size, which is around 8 times smaller in a typical quantitative shotgun proteomics experiment compared to state-of-the-art microarray data. However, testing for the hypothesis by under-sampling microarray datasets suggests that this is unlikely to be the case (Fig. 38) and that there has to be a more fundamental reason as to why different miRNA apparently regulate different targets at the mRNA and at the protein levels.

We therefore sought to understand the system better by taking a closer look at the properties of functional miRNA binding sites in proteomics experiments. As, miRNAs repress translation and turn up

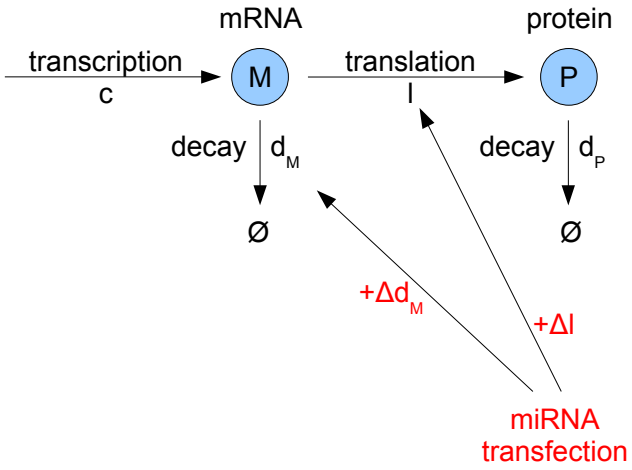


Figure 20: A six parameters – two state variables model of gene expression regulation by miRNAs. M and P represent the mRNA and protein concentration of a single gene respectively. Four parameters c, l, d_M and d_P describe the transcription, translation, mRNA decay and protein decay rates. The transcription and protein decay rates c and d_P are assumed to be left unchanged by a change in the miRNA level, while the mRNA decay and translation rates d_M and l are changed by an amount Δd_M and Δl. Such a model was previously introduced by Khanin and Vinciotti [120].

mRNA decay, changes in protein levels are influenced by these two mechanisms of action. Microarray data, on the other hand, focus on the effect of miRNAs on mRNA decay.

Following a similar approach to Figure 8, where we proposed that the binding of miRNAs to target sites depends mostly on structural properties of the miRNA binding site while subsequent mRNA degradation depends on sequence properties, we set to isolate the specific effect of translation repression from measurements of changes in protein and mRNA levels, and subsequently, to characterize the properties of miRNA binding sites that specifically cause translation repression.

To perform this analysis, we first developed a method to estimate the changes in translation rates induced by a transfected miRNA from the measured changes in protein and mRNAs levels.

6.2 A SIMPLE MODEL TO ESTIMATE MIRNA-INDUCED CHANGES IN TRANSLATION RATES

Figure 20 presents a coarse-grained ordinary differential equation model of gene regulation by a miRNA. In the absence of the miRNA, this models defines the dynamics of the concentration of one protein P encoded by one mRNA M as:

$$\begin{cases} \frac{dM}{dt} &= c - d_M M \\ \frac{dP}{dt} &= lM - d_P P \end{cases}$$

The steady state of that system is:

$$\begin{cases} M^* &= \frac{c}{d_M} \\ P^* &= \frac{lc}{d_M d_P} \end{cases}$$

Setting $M(0) = M_0$ and $P(0) = P_0$, we can solve the system:

$$\begin{cases} M(t) &= \frac{c}{d_M} + \left(M_0 - \frac{c}{d_M}\right) e^{-d_M t} \\ P(t) &= \frac{lc}{d_P d_M} + \frac{l}{d_P - d_M}\left(M_0 - \frac{c}{d_M}\right) e^{-d_M t} + \left(P_0 - \frac{l}{d_P - d_M}\left(M_0 - \frac{c}{d_P}\right)\right) e^{-d_P t} \end{cases}$$

Let us consider that at the beginning of the experiment $t < 0$, the system is at steady state:

$$\forall t < 0,\ (M(t), P(t)) = (M_0, P_0) = \left(\frac{c}{d_M}, \frac{lc}{d_M d_P}\right)$$

At time $t = 0$, the miRNA is transfected, which has the effect of changing d_M and l (see figure 20):

$$\begin{cases} d_M &\mapsto d_M + \Delta d_M & \text{with } \Delta d_M \geqslant 0 \\ l &\mapsto l + \Delta l & \text{with } \Delta l \leqslant 0 \end{cases}$$

From $t = 0$ on, the system follows the new dynamics and evolves towards the new steady state

$$(M(t_\infty), P(t_\infty)) = \left(\frac{c}{d_M + \Delta d_M}, \frac{(l + \Delta l)c}{(d_M + \Delta d_M) d_P}\right)$$

The microarray and shotgun proteomics experiments measure two different log expression ratios.

From the microarray experiments, we obtain:

$$\log\left(\frac{M(t_\infty)}{M(0)}\right) = \log\left(\frac{d_M}{d_M + \Delta d_M}\right)$$

In other words, what the microarray really measures at steady-state is the miRNA-induced *relative* change in mRNA-decay rate. For instance, if the $\log_2$ fold change in mRNA level following the miRNA transfection is -1:

$$\begin{aligned} \log_2\left(\tfrac{M(t_\infty)}{M(0)}\right) &= \log_2\left(\tfrac{d_M}{d_M + \Delta d_M}\right) &= -1 \\ &\Leftrightarrow \Delta d_M &= d_M \end{aligned}$$

That is, the miRNA caused the mRNA decay rate d_M to double.

Similarly, the shotgun proteomics measures the log ratio:

$$\begin{aligned} \log\left(\tfrac{P(t_\infty)}{P(0)}\right) &= \log\left(\tfrac{l+\Delta l}{l}\tfrac{d_M}{d_M+\Delta d_M}\right) \\ &= \log\left(\tfrac{l+\Delta l}{l}\right) + \log\left(\tfrac{M(t_\infty)}{M(0)}\right) \end{aligned}$$

and so,

$$\log\left(\frac{P(t_\infty)}{P(0)}\right) - \log\left(\frac{M(t_\infty)}{M(0)}\right) = \log\left(1 + \frac{\Delta l}{l}\right) \tag{6.1}$$

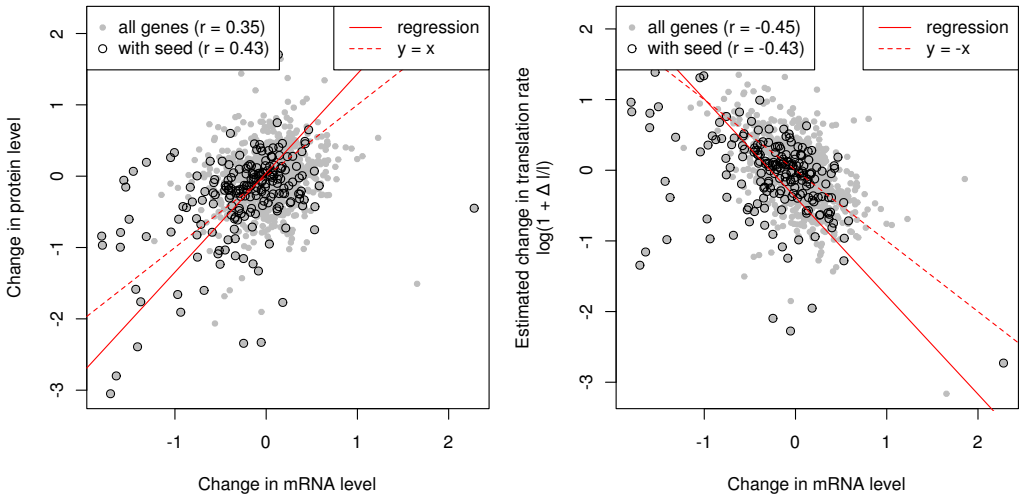


Figure 21: Effect of miR-124 transfection on mRNA and protein levels in a SILAC experiment [8]. Left: correlating changes in mRNA level with changes in the cognate protein levels for all mRNAs ($r = .35$) or mRNAs with seed match ($r = .43$). The full red line represent the first principal component of the scatter of all mRNAs with seed match, while the dotted red line represents the $y = x$ line. Right: correlating changes in mRNA level with estimated changes in the translation rate — *i.e.* $\log\left(1 + \frac{\Delta l}{l}\right)$ — under the model described in section 6.2, for all mRNAs ($r = -.45$) or all mRNAs with seed match ($r = -.43$). The full red line represents the first principal component of the scatter of all mRNAs with seed match, while the dotted red line represents the $y = -x$ line.

In other words, *assuming the measurements are performed at the new steady state* ($t = t_\infty$), subtracting the change at the protein level from the change at the mRNA level estimates the miRNA-induced relative change in the translation rate. In addition, we see that in this model, observing that the changes in mRNA level differ from changes in protein level is evidence for miRNA-induced regulation of the mRNA translation rate (l).

6.2.1 *Application to SILAC proteomics and transcriptomics data*

The left panel of Figure 21 puts the model we just introduced in the context of real proteomics and transcriptomics data, namely the joint Stable Isotope Labeling with Aminoacids in cell Culture (SILAC) proteomics and transcriptomics measurements following miRNA transfection of Baek et al. [8]. Such experiments measure changes in protein levels 48h after miRNA transfection in HeLa cells, together with the miRNA induced changes in mRNA levels 24h after miRNA transfection. As previously observed by Baek et al. [8], genes with seed match to the transfected miRNA show decreased mRNA and protein levels, which induces a positive correlation between changes in mRNA and protein levels. These observations are consistent with miRNAs repressing gene expression both at the protein and mRNA levels. The correlation is slightly higher for genes with with seed matches ($r = .43$) than for all measured genes ($r = .35$), consistent with a direct effect of miR-

NAs on the mRNA and protein levels of these target genes. Finally, for genes with seed match, comparing the regression line (full red) to the $y = x$ line (dotted red), we see that changes in protein levels are on average larger than changes in mRNA levels. Under the model's assumptions, this would reflect the fact that changes in protein levels reflect miRNA-induced changes in the translation rate on top of the effect at the mRNA level.

However, a closer look at the effect of the transfected miRNA on the translation and degradation of the target mRNAs suggests that the data is more complex.

The right panel of Figure 21 correlates changes in mRNA level to estimated changes in translation rate upon miRNA transfection. Changes in translation rates were estimated using the model introduced in section 6.2, by subtracting changes in mRNA levels from changes in protein levels (equation 6.1). Because miRNAs have been shown to repress translation and increase mRNA decay, one would expect changes in translation rates to correlate positively with changes in mRNA levels of miRNA target genes. On the other hand, we do not expect any correlation between the changes in translation rate and the changes in mRNA levels of genes that are not targeted by the miRNA, as those should not be affected by the miRNA transfection. However, the right panel of Figure 21 is in total disagreement with these expectations: estimated changes in translation rates correlate negatively with changes in mRNA levels, as if the larger the drop in mRNA concentration, the stronger the increase in translation rate. In addition, the correlation is unexpectedly similar for genes with seed match ($r = -.45$) compared to all genes ($r = -.43$). This would imply that when it comes to translation regulation, putative miRNA target genes are no different than other genes.

6.2.2 *Application to pulsed-SILAC proteomics and transcriptomics data*

To gain more confidence in these observations, we repeated the analysis using the proteomics and transcriptomics data from Selbach et al. [191], who performed pulsed Stable Isotope Labeling with Aminoacids in cell Culture (pSILAC). Very briefly, the experimental procedure differs from the SILAC experiments of Baek et al. [8] in that it measures changes in the amounts of newly synthesized protein between 8h and 32h after miRNA transfection in HEK293 cells. In parallel, mRNA levels were profiled by microarrays at 0h, 8h and 32h after miRNA transfection. From these measurements, we obtained changes in mRNA levels 32h after miRNA transfection.

The left panel of Figure 22 reveals a pattern similar to that of Figure 21. Putative target genes of the transfected miRNA show reduced mRNA and protein levels, resulting in a positive correlation between changes in mRNA and protein levels. In addition, this correlation is higher for genes with seed match ($r = .43$) compared to all measured genes ($r = .19$), possibly highlighting the direct regulatory effects of the transfected miRNAs. However, contrary to Figure 21, changes in protein levels are not larger than changes in mRNA levels on average, which is inconsistent with the prediction of our model. In addition, turning the right panel of Figure 22, we observe a strong negative correlation between the estimated changes in translation rate and the changes in mRNA level instead of the expected positive correla-

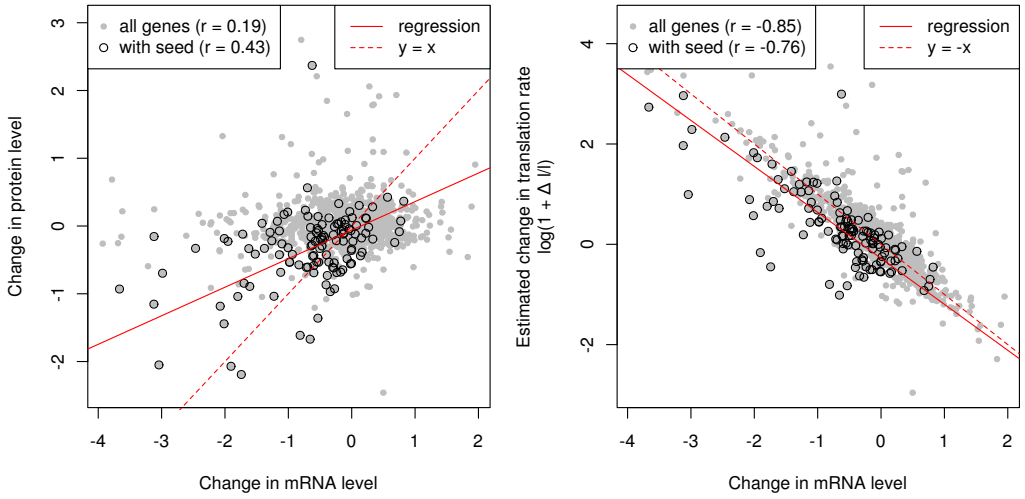


Figure 22: Effect of miR-155 transfection on the mRNA and protein levels in a pSILAC experiment [191]. Left: correlating changes in mRNA level with changes in the cognate protein levels for all mRNAs ($r = .19$) or mRNAs with seed match ($r = .43$). The full red line represent the first principal component of the scatter of all mRNAs with seed match, while the dotted red line represents the $y = x$ line. Right: correlating changes in mRNA level with estimated changes in the translation rate — *i.e.* $\log\left(1 + \frac{\Delta l}{l}\right)$ — under the model described in section 6.2, for all mRNAs ($r = -.85$) or all mRNAs with seed match ($r = -.76$). The full red line represents the first principal component of the scatter of all mRNAs with seed match, while the dotted red line represents the $y = -x$ line.

tion. This applies both to genes with seed matches and to the general population of measured genes, similar to the right panel of Figure 21. These observations also hold if we use matched time points, comparing changes in mRNA levels between 8h and 32h after miRNA transfection (not shown).

In conclusion, under the model presented in this section, changes in protein and mRNA levels after miRNA transfection from both SILAC and pSILAC experiments show fundamental inconsistencies with the expectations of current miRNA biology. This then reflects on properties of functional miRNA binding sites that we inferred from these datasets in Chapter 3 (see Figure 5). As this model is merely an attempt to represent how miRNAs interfere with gene expression in the most simple way, it appears important to propose an explanation to the inconsistencies observed so far from the different analyses we have performed.

6.3 MIRNA INDUCED CHANGES IN GENE EXPRESSION ARE FAR FROM STEADY-STATE IN STATE-OF-THE-ART PROTEOMICS MEASUREMENTS

One possible explanation to these inconsistencies is prompted by the observation that a minority of genes with seed match do appear to have reduced translation and lower mRNA levels in response to the miRNA transfection (Figures 21 and 22). It could therefore be that most mRNAs harboring a seed match to the transfected miRNA are actually not functional targets of the transfected miRNA. In other words, they behave like genes without seed match and therefore, for most of them, we expect changes in translation rates to be independent from changes in mRNA decay rates.

Another explanation stems from the observation in the biochemical literature that miRNAs trigger a rapid inhibition of translation initiation, followed by slower deadenylation and decapping, which ultimately lead to increased mRNA degradation [66, 62]. Could it be that the time-scales of these two types of regulation have a strong influence on the mRNA and protein levels of target genes in these experiments, and that the precise kinetics of miRNA regulation play a major role in proteomics measurements? Doherty et al. [48] measured that the half-life of a number of 40S ribosomal subunit proteins are in the 25h-30h range, which is about the length of a typical proteomics experiment. Therefore, in a miRNA transfection experiment, fast-decaying protein A is expected to display stronger down-regulation than slow-decaying protein B, even if the mRNAs of A and B are equally well bound by the transfected miRNA. In addition, for protein with long half-lives, the miRNA-induced down-regulation of the target mRNA levels will only reflect on the protein levels after a significant delay. Such effects could result in an apparent decoupling of the changes in protein level from the changes in mRNA level.

To test these hypotheses, we will now set up a probabilistic model M_0 on top of the ordinary differential equation model of section 6.2. From the proteomics and transcripts measurements in the presence and absence of the miRNA, we can obtain two vectors $f_P = \log_2\left(\frac{P(t_\infty)}{P(0)}\right)$ and $f_M = \log_2\left(\frac{M(t_\infty)}{M(0)}\right)$ representing the log2 fold change of protein and mRNA levels following miRNA transfection.

As previously mentioned in chapters 1 and 3, the best single determinant of miRNA targeting is the presence of a miRNA seed match in the 3'UTR of a mRNA. Therefore, of all mRNAs and proteins present in the dataset, we will focus on those that have a seed match. But a mRNA harboring a seed match in its 3'UTR is not necessarily a functional target of the cognate miRNA. Therefore, we assume that the population of mRNAs with seed match can be split into two sub-categories: mRNAs carrying functional seed matches (+), and mRNAs carrying non-functional seed matches (−). Under these assumptions, we will now write a probabilistic model of high-throughput proteomics and transcriptomics measurements following miRNA transfection. Depending on the i^{th} mRNA/protein being a functional (+) or a non-functional (−) miRNA target, we can write the probability of measuring a change $f_{M,i}$ at the mRNA level and $f_{P,i}$ at the protein level as:

$$\begin{cases} P(f_{M,i}|+) &= \mathcal{N}\left(\mu_{M,+}, \sigma^2_{M,+}\right) \\ P(f_{P,i}|+) &= \mathcal{N}\left(f_{M,i} + \mu_{P,+}, \sigma^2_{P,+}\right) \\ P(f_{M,i}|-) &= \mathcal{N}\left(\mu_{M,-}, \sigma^2_{M,-}\right) \\ P(f_{P,i}|-) &= \mathcal{N}\left(f_{M,i} + \mu_{P,-}, \sigma^2_{P,-}\right) \end{cases} \tag{6.2}$$

Here, we explicitly assume that the log2 fold changes of target mRNAs are Gaussian-distributed around a mean value $\mu_{M,+}$ which captures the average effect of miRNAs on mRNA decay. More precisely, $\mu_{M,+}$ is the log2 relative change in mRNA decay induced by the miRNA, averaged over all target mRNAs, as established in section 6.2. $\sigma_{M,+}$ represents the variability of the effect of miRNAs on mRNA decay across all target mRNAs. $\mu_{P,+}$ is the miRNA-induced average change in the translation rate of target mRNAs, while $\sigma_{P,+}$ represents the variability on the changes in protein levels across target mRNAs. We see that the model implies that $f_P = f_M + \mu_P$ on average, that is `protein change = mRNA change + translation change`, consistent with equation (6.1).

In addition to these four parameters, we also introduce $\mu_{M,-}$, $\mu_{P,-}$, $\sigma_{M,-}$, $\sigma_{P,-}$ to capture the secondary effects of the miRNA transfection on mRNAs and protein that are not functional targets of the transfected miRNA. Finally, we define ρ as the (yet unknown) fraction of functional targets among mRNAs with a seed match to the transfected miRNA:

$$\rho := P(+)$$

This probabilistic model is very similar to the one we used to determine functional miRNA binding sites from replicated miRNA transfection and microarray experiments in chapter 3, detailed in section B.1.2. The major difference is that in the present section, we explicitly model how miRNAs interfere with gene expression at the protein and mRNA levels and do not make use of replicates since those are not available in the two datasets we are analyzing here [8, 191].

6.3.1 *Estimating the parameters*

We estimate $\mu_{M,-}$, $\mu_{P,-}$, $\sigma_{M,-}$, $\sigma_{P,-}$ using the maximum likelihood estimator of the mean and variance from mRNAs and proteins that

surely are not miRNA targets, namely all mRNAs (and cognate proteins) that do not harbor a seed match to the transfected miRNA.

The model presented here is a two-component Gaussian mixture model, where we just fixed one mixture using measurements from mRNAs and proteins that are not being modeled here because they do not harbor a seed match. We can write down the posterior probability of being a functional miRNA target site given the data (f_P, f_M):

$$\begin{aligned} P(+|f_M, f_P) &= \frac{P(f_M, f_P|+)P(+)}{P(f_P, f_M)} \\ &= \frac{P(f_P|f_M, +)P(f_M|+)P(+)}{P(f_P|f_M)P(f_M)} \\ &= \frac{P(f_P|f_M, +)P(f_M|+)\rho}{P(f_P|f_M, +)P(f_M|+)\rho + P(f_P|f_M, -)P(f_M|-)(1-\rho)} \end{aligned}$$

We will now estimate the remaining parameters $\theta = (\sigma_{M,+}, \sigma_{P,+}, \mu_{M,+}, \mu_{P,+}, \rho)$ by Expectation Maximization (EM) [44]. The log-likelihood of these parameters is:

$$\begin{cases} \mathcal{L}(\theta|f_{P,i}, f_{M,i}, +) &= \log\rho - \log\sigma_{P,+} - \log\sigma_{M,+} - \log 2\pi \\ & \quad -\frac{1}{2}\left(\frac{f_{P,i}-(f_{M,i}+\mu_{P,+})}{\sigma_{P,+}}\right)^2 - \frac{1}{2}\left(\frac{f_{M,i}-\mu_{M,+}}{\sigma_{M,+}}\right)^2 \\ \mathcal{L}(\theta|f_{P,i}, f_{M,i}, -) &= \log(1-\rho) - \log\sigma_{P,-} - \log\sigma_{M,-} - \log 2\pi \\ & \quad -\frac{1}{2}\left(\frac{f_{P,i}-(f_{M,i}+\mu_{P,-})}{\sigma_{P,-}}\right)^2 - \frac{1}{2}\left(\frac{f_{M,i}-\mu_{M,-}}{\sigma_{M,-}}\right)^2 \end{cases}$$

We can then write down the E-step of the expectation maximization algorithm as:

$$\begin{cases} X_{i,+}^{(t)} &= P\left(+|f_{P,i}, f_{M,i}, \theta^{(t)}\right) \\ X_{i,-}^{(t)} &= P\left(-|f_{P,i}, f_{M,i}, \theta^{(t)}\right) \end{cases}$$

and the M-step:

$$\rho^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} X_{i,+}^{(t)}$$

$$\begin{aligned} \mu_{M,+}^{(t+1)} &= \frac{\sum_{i=1}^{n} X_{i,+}^{(t)} f_{M,i}}{\sum_{i=1}^{n} X_{i,+}^{(t)}} \\ \sigma_{M,+}^{(t+1)} &= \sqrt{\frac{\sum_{i=1}^{n} X_{i,+}^{(t)}\left(f_{M,i}-\mu_{M,+}^{(t+1)}\right)^2}{\sum_{i=1}^{n} X_{i,+}^{(t)}}} \\ \mu_{P,+}^{(t+1)} &= \frac{\sum_{i=1}^{n} X_{i,+}^{(t)}\left(f_{P,i}-f_{M,i}\right)}{\sum_{i=1}^{n} X_{i,+}^{(t)}} \\ \sigma_{P,+}^{(t+1)} &= \sqrt{\frac{\sum_{i=1}^{n} X_{i,+}^{(t)}\left(f_{P,i}-f_{M,i}-\mu_{P,+}^{(t+1)}\right)^2}{\sum_{i=1}^{n} X_{i,+}^{(t)}}} \end{aligned}$$

We then iterate between the E- and the M- step until the log-likelihood $\mathcal{L}(\theta|f_{P,i}, f_{M,i})$ converges.

6.3.2 *An alternative model*

Finally, we introduce an alternative model $M_\perp$ of (p)SILAC and miRNA transfection experiments, in which changes in protein level depend

Dataset	Model	rho	$\mu_{M,+}$	$\sigma_{M,+}$	$\mu_{P,+}$	$\sigma_{P,+}$	$\mathcal{L}(\theta^*\|f_M, f_P)$
Baek miR-124	M_0	0.28	-0.69	0.73	0.02	0.99	-293.55
Baek miR-124	$M_\perp$	0.30	-0.68	0.69	-0.78	0.82	-275.55
Selbach miR-155	M_0	0.17	-1.93	0.84	1.45	0.80	-268.97
Selbach miR-155	$M_\perp$	0.40	-1.18	0.94	-0.62	0.72	-237.23

Table 1: Best-fitted parameters of miRNA regulation under models M_0 and $M_\perp$ from the miR-124 transfection data of Baek et al. [8] and the miR-155 transfection data of Selbach et al. [191].

only on the action of miRNA on the translation rate, not on changes in the cognate mRNA levels:

$$\begin{cases} P(f_{M,i}|+) &= \mathcal{N}\left(\mu_{M,+}, \sigma^2_{M,+}\right) \\ P(f_{P,i}|+) &= \mathcal{N}\left(\mu_{P,+}, \sigma^2_{P,+}\right) \\ P(f_{M,i}|-) &= \mathcal{N}\left(\mu_{M,-}, \sigma^2_{M,-}\right) \\ P(f_{P,i}|-) &= \mathcal{N}\left(\mu_{P,-}, \sigma^2_{P,-}\right) \end{cases} \tag{6.3}$$

Like with model M_0, we use the EM algorithm to fit the parameters. To do so requires minor changes in the E- and M-steps.

So model $M_\perp$ defined by equation (6.3) assumes that changes in protein levels are independent from changes at the mRNA level, while model M_0 defined by equation (6.2) assumes changes in protein levels reflect the action of miRNAs on the translation rate as well as the miRNA-induced changes in mRNA level, in agreement with the ordinary differential equation model of section 6.2 and equation (6.1).

6.3.3 *The parameters obtained under M_0 are inconsistent with the expectations of miRNA biology*

Table 1 shows the best-fitted parameters for models M_0 and $M_\perp$ using the miR-124 transfection data from Baek et al. [8]. We observe that under the assumptions of both M_0 and $M_\perp$, an estimated $\rho \simeq 30\%$ of genes with seed matches are inferred to be functional target of the transfected miRNA. The mRNA of these target genes appear to go down by $\mu_{M,+} = -.7$ on log2 scale, which corresponds to mRNAs dropping to 60% of their original levels in response to the miRNA. The estimate of the standard deviation $\sigma_{M,+}$ is comparable to the average miRNA effect on mRNA stability, suggesting that the effect of miRNAs on mRNA levels is quite "noisy", with a lot of inter-gene variability. However, differences between M_0 and $M_\perp$ stand out dramatically when looking at parameters related to translation regulation. Under M_0, miRNAs appear to have virtually no effect on translation ($\mu_{P,+} = .02$), as if changes in protein levels corresponded to changes in mRNA levels with large amounts of noise ($\sigma_{P,+} = .99$). $M_\perp$ gives a different picture, where miRNAs reduce protein levels of target genes by $\mu_{P,+} = -.78$, which corresponds to proteins dropping to 60% of their original levels.

As the magnitude of the regulatory effect of miRNAs is comparable at the protein level and at the mRNA level, it may be tempting

to interpret these observations by proposing that miRNA only act by decreasing mRNA levels, with changes in cognate proteins levels subsequently reflecting the decreased mRNA levels. However, this is not what model comparison suggests: given the measure of the goodness of fit in column $\mathcal{L}(\theta^*|f_M, f_P)$, we will show in section 6.3.4 that $M_\perp$ fits the data much better M_0. This leads to a different interpretation: protein and mRNA levels may be changing by the same amount on average, but these changes are not coordinated the way they should be if mRNAs levels were reflecting on protein levels.

Before elaborating further on model comparison and possible biological interpretations, we will look at the parameters inferred from the miR-155 transfection experiment of Selbach et al. [191], shown in the 3rd and 4th row of Table 1. There, the only common feature seems to be the amount of gene-to-gene variability ($\sigma_{M,+}, \sigma_{P,+}$), while the remaining parameters estimates differ quite significantly. Most striking is the specific effect of miRNAs on protein levels which is positive under M_0 ($\mu_{P,+} = 1.45$) which would suggest a 2.7-fold increase in the translation rate of miRNA targets. This is clearly inconsistent with the estimate under $M_\perp$ ($\mu_{P,+} = -.62$) and the established paradigm in which miRNAs act as translational repressors.

In summary, it seems like using M_0 to estimate the effect of miRNA regulation on protein and mRNA levels of target genes leads to results that are inconsistent with the findings of miRNA biology.

We will now examine which of the two models M_0 or $M_\perp$ is best supported by the data.

6.3.4 *SILAC and pSILAC experiments support a model in which changes in protein and mRNA levels are decoupled*

To determine which of model M_0 or $M_\perp$ agrees most with the data (f_M, f_P), we will compute the odds ratio between the two models given the data. Using Bayes' theorem and equal *a priori* probability of the two models $P(M_0) = P(M_\perp)$, we get:

$$\begin{aligned} \log_{10}\left(\frac{P(M_0|f_M,f_P)}{P(M_\perp|f_M,f_P)}\right) &= \log_{10}\left(\frac{P(f_M,f_P|M_0)P(M_0)}{P(f_M,f_P)}\frac{P(f_M,f_P)}{P(f_M,f_P|M_\perp)P(M_\perp)}\right) \\ &= \log_{10}\left(\frac{P(f_M,f_P|M_0)}{P(f_M,f_P|M_\perp)}\right) \end{aligned} \tag{6.4}$$

To compute this odds ratio, we need to integrate the probability of the data under each model, weighted by the a priori probability of the parameters:

$$P(f_M, f_P|M_0) = \int P(f_M, f_P|\theta, M_0)P(\theta)d\theta$$

which can be in principle be evaluated by multi-dimensional numerical integration. Alternatively, we can approximate $P(\theta)$ by a Dirac delta function:

$$P(\theta) \simeq \begin{cases} 1 & \text{if } \theta = \theta^* \\ 0 & \text{if } \theta \neq \theta^* \end{cases}$$

where θ^* are the parameters inferred by the EM algorithm. Plugging this approximation in the integral gives

$$P(f_M, f_P|M_0) \simeq P(f_M, f_P|\theta^*, M_0)$$

where $P(f_M, f_P|\theta^*, M_0)$ is the probability of the data under the likeliest parameters, whose logarithm is reported in the column $\mathcal{L}(\theta^*|f_M, f_P)$ of Table 1. To summarize:

$$\log_{10}\left(\frac{P(M_0|f_M, f_P)}{P(M_\perp|f_M, f_P)}\right) \simeq \frac{1}{\log 10}\left(\mathcal{L}(\theta^*|f_M, f_P, M_0) - \mathcal{L}(\theta^*|f_M, f_P, M_\perp)\right)$$

In the miR-124 experiment of Baek et al. [8], the odds are $1 : 10^8$ in favor of $M_\perp$. The same applies to the miR-155 experiment of Selbach et al. [191] where the odds are $1 : 10^{14}$ in favor of $M_\perp$. This can be confirmed visually by sampling proteomics and transcriptomics datasets under the best-fitted parameters and then comparing the datasets "expected" by M_0 and $M_\perp$. As Figure 23 shows, the datasets generated by M_0 and $M_\perp$ look very different qualitatively. Comparing the top row of Figure 23, it is clear that the data simulated under $M_\perp$ reproduces the data of Baek et al. [8] shown on Figure 21 best. This can be seen from the shape of the scatters of all genes vs genes with seed match, as well as from the correlation coefficients and the position of the first principal component relative to the $y = x$ line. Turning to the data from Selbach et al. [191], it is also clear that the data simulated from $M_\perp$ reproduces Figure 22 best.

Therefore, we can conclude from our analysis that in the proteomics data of Baek et al. [8] and Selbach et al. [191], there is no strong coupling between changes in mRNA levels and changes in protein levels. In other words, in these datasets, protein and mRNAs levels are changing for different reasons. While the mRNA and protein levels of genes targeted by the transfected miRNA both go down in these experiments, it is also clear that changes in mRNA levels are not reflected in the changes in protein levels. Of course, we expect that changes in mRNA would ultimately propagate to protein levels. But this appears not to be the case at the time-points at which the measurements were performed, which suggests that the mRNA and protein levels are far from the steady-state we assumed in section 6.2, and that the steady-state ODE model we introduced cannot be used to interpret the proteomics data of Baek et al. [8] and Selbach et al. [191]. Instead, our results suggest that interpreting changes in protein levels in such experiments requires explicit modeling of the transient behavior of mRNA and protein levels.

6.4 A DETAILED ODE MODEL OF MIRNA-MEDIATED GENE REGULATION

6.4.1 *Questions we would like to address*

There are three concrete motivations to modeling the kinetics of miRNA-mediated gene regulation.

First of all, such a model could be used to provide a final answer to the interrogations prompted in Chapter 3 as to why miRNA binding sites that lead to down-regulated protein levels in a miRNA transfection experiment do not share the properties of functional miRNA binding sites. To do so, one could first show that the statistical overrepresentation of the miRNA seed motif within the 3'UTR of mRNAs

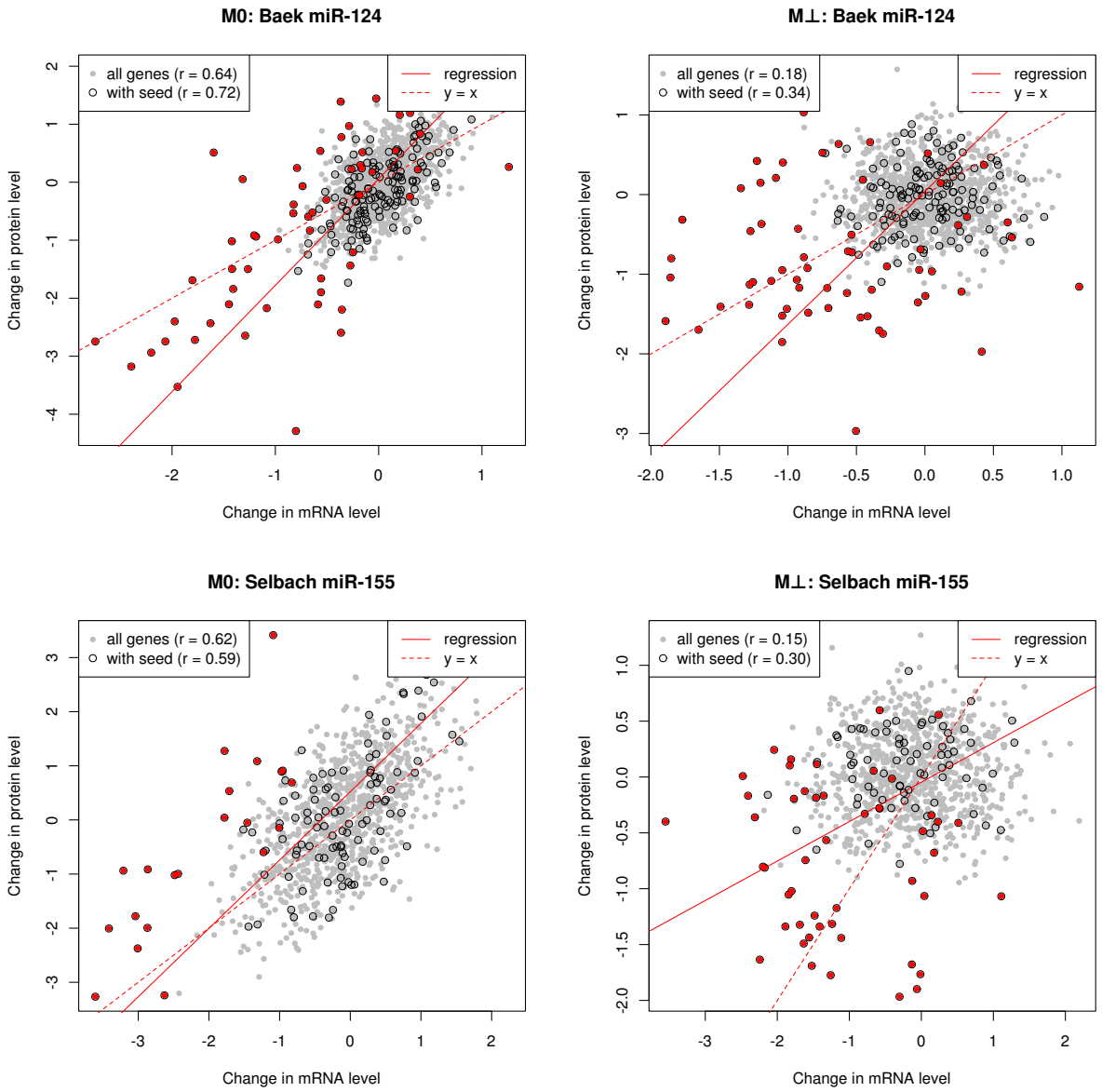


Figure 23: Sampling proteomics and transcriptomics datasets from M_0 (left column) and $M_\perp$ (right column) using parameters best-fitted to the miR-124 experiment of Baek et al. [8] (top row) and the miR-155 experiment of Selbach et al. [191]. Shown are scatters of the simulated changes in protein and mRNA levels for all genes (grey dots), genes with seed match (black circles), and genes with functional seed match (red dots). As in Figures 21 and 22, we report the correlation coefficients between changes in protein and mRNA levels for all genes, or genes with a seed match to the transfected miRNA. The full red line represents the first principal component of the scatter for genes with seed match.

coding for down-regulated proteins improves if one uses the model to "correct" the measured changes in protein and mRNA levels [191, 8] for gene-dependent parameters such as mRNA and protein turn-over, or the affinity of the miRNA to the target mRNA. In case this can be proven, one would then proceed to show that miRNA binding sites leading to down-regulated protein levels in a miRNA transfection experiment have the expected properties of functional miRNA binding sites (accessibility, sequence composition, etc.) if we use the model-"corrected" changes in protein level instead of the measured changes.

Second, one could use the model to suggest improvements to high-throughput experiments aimed at identifying functional miRNA target genes by means of perturbing miRNA expression. Based on the model predictions of changes in mRNA and protein levels over time, one would compute an optimal experimental time-point for the genes of interest, defined as the amount of time after miRNA perturbation after which changes in the mRNA and protein levels of predicted target genes are expected to be largest. One would then perform a miRNA transfection experiment and measure changes in gene expression at the optimal time-point as well as at a standard time-point (such as 48h after transfection). From such data, one would attempt to show that the data gathered at the optimal time-point is statistically better behaved than the standard time-point, for instance by comparing the over-representation of the miRNA seed motif in differentially-regulated genes at the two experimental time points.

Finally, one could investigate open questions regarding the mechanism of miRNA action. One such question is whether miRNAs repress mRNA translation in addition to increasing mRNA decay. A large body of experimental work suggests that this is the case [231, 211, 66, 62, 240, 64, 46, 63]. The major strength of such studies is that they make it possible to obtain a detailed, reductionist view on the mechanism of miRNA action. But this comes at the expense of working with artifical reporter systems that display a strong response to miRNAs, or with well-controlled minimal cell-free set-ups which may or may not accurately model miRNA regulation *in vivo*. In addition, such studies can only examine a limited amount of miRNA target genes or miRNA target sites, giving only little insight on miRNA regulation at the genome scale. For instance, maybe miRNAs increase the decay rate of all target mRNAs while only repressing the translation of a minority of them. Such hypothesis can only be addressed by high-throughput studies, such as those performed by Hendrickson et al. [99] and Guo et al. [88], which lead to conflicting results: Hendrickson et al. [99] observed that genes targeted by miRNAs show concordant changes in mRNA levels and translation rate, while Guo et al. [88] proposed that the effect of miRNAs on protein levels can be explained by changes in mRNA levels, that is miRNA do not repress translation. Another open question is whether RISC complexes recycle at 100% after the target mRNA is degraded, or if some RISC is also lost in the process. Using our model and an approach similar to the one we use in section 6.3, we could look at these questions by making competing models of miRNA action with different topologies reflecting the alternative hypotheses, fit the models to the data, and determine which is likelier given high-throughput [8, 191, 99, 88] or detailed, small scale data.

STATE OF THE FIELD To answer these three questions requires an accurate model of the kinetics of miRNA action. Several studies have proposed such models, yet with goals different from the ones we set to ourselves here.

Bartlett and Davis [14] developed a model of siRNA action which cannot be used our purpose as it focuses on siRNA-mediated mRNA cleavage and does not take miRNA-induced translation repression into account.

Khanin and Vinciotti [120] introduces an ODE model of miRNA action which takes time-varying miRNA levels into account but does not model the evolution of protein levels over time. Khanin and Higham [119] introduces a model of miRNA action at the mRNA and protein levels while Maya [148] looks at a 2- and 3-step stochastic version of the same model, including one miRNA – 2 targets models. These two models assume that miRNAs have an instant and simultaneous effect on the translation and decay rates of target mRNAs. This may or may not be appropriate when it comes to modeling experimental measurements, where the transfected miRNA first needs to compete with endogenous miRNAs to load into an Ago protein. The effect of that competition — a slight up-regulation of the endogenous miRNA target genes — are clearly observable in high-throughput studies [118], which suggests that these effects and the induced delay in gene regulation are important and need to be taken into account. In addition, it is not clear whether one can make the assumption that miRNAs inhibit translation and increase mRNA decay on the same time scale [66, 62]. Therefore the assumption that miRNAs have a simultaneous effect on the translation and decay rates of target mRNAs may not hold when it comes to modeling the kinetics of miRNA action.

Other work on modeling miRNA-mediated gene regulation include Levine et al. [133], which proposes a model of small RNA gene silencing in bacteria, and Zhdanov [239] that examines the requirements for efficient mRNA repression by miRNAs. Finally, Stanhope et al. [202] introduces a regression model to characterize the relationship between miRNAs levels and target mRNA levels, taking factors such the RISC occupancy into account. However, none of these study explicitly take the timing of miRNA action into account.

Maybe the closest work to the one we envision here is Wang et al. [225], which models the kinetics of miRNA action using ODE and stochastic simulations to identify key steps in the miRNA regulation pathway and quantify stochastic noise strength along the pathway. A model is introduced, that has a topology similar to the one we propose in section 6.4.2, but with a purpose different than ours, which is reflected by two important differences: the competition of the exogenous miRNAs with the preexisting endogenous miRNAs is not modeled, and miRNAs are assumed only to mediate translation repression while leaving the mRNA decay rate unchanged. In addition, about a third of the reactions rates in that model are presently unknown because they were never measured. Wang et al. [225] addressed this problem by plugging in biologically plausible values into the model. But to go beyond a qualitative description of the kinetics of miRNA regulation, one probably needs to obtain data-fitted estimates of the reaction-rates. Finally, the study does not provide a measure of how accurately the model can reproduce the data. In our work, we aim at addressing these three points specifically.

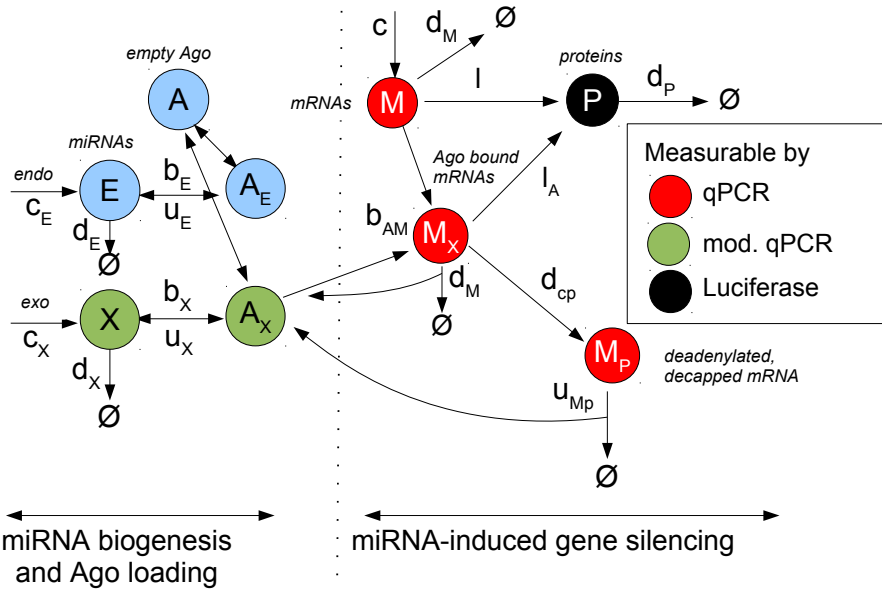


Figure 24: A 17 parameters – 9 state variables ordinary differential equation model of miRNA-mediated gene regulation. Not shown on the figure are R, the total amount of RISC/Ago, and the inducible promoter leakage rate. Details are in the text. States are described in Table 2 and reaction rates in Table 3.

6.4.2 *Model structure*

We use an ordinary differential equation model to describe how one miRNA regulates the mRNA and protein levels of a single target gene. The topology of the proposed model illustrated on figure 24 attempts to summarize the mechanism of miRNA action established by genetics and biochemistry [62, 66]. The state variables are described in Table 2 and the reaction rates in Table 3. Briefly, in the absence of exogenous miRNAs ($X = 0$), endogenous miRNA (E) are synthesized (c_E), decay (d_E), load into a free Ago A with rate $b_E AE$ or unload from a loaded Ago A_E with rate $u_E A_E$. These miRNAs do not target the mRNA M that codes for a fluorescent protein P, which decays at the rate $d_P P$. mRNAs are produced at constant rate c, decay at rate $d_M M$ and are translated at rate Ml.

At time $t = 0$, we induce the transcription of an exogenous miRNA X. Upon induction, mature miRNAs will be produced at rate c_X and decay at rate $d_X X$. This miRNA needs to compete for free Ago with the preexisting endogenous miRNAs before it can load into a free Ago with rate $b_X AX$. Once the RISC complex A_X is formed, it can either dissociate with rate $u_X A_X$ or bind a free target mRNA M to form a ternary Ago-miRNA-mRNA complex M_X with rate $b_{AM} A_X M$. As a result, the translation rate of the bound mRNA drops from l to l_A per mRNA copy. The ternary complex will then recruit the GW182/TNRC6 protein which will ultimately lead to deadenylation and decapping of the target mRNA at rate $d_{cp} M_X$. The deadenylated, decapped mRNA M_P cannot be translated anymore and its decay rate increases from

State	Description
E	free endogenous miRNAs
X	free exogenous miRNA
A	free Ago
A_E	Ago-bound endogenous miRNAs
A_X	Ago-bound exogenous miRNA
M	free target mRNA
M_X	Ago-miRNA-mRNA ternary complex
M_P	deadenylated, decapped mRNA
P	target protein

Table 2: The 9 state variables in the model.

Reaction rate	Description
c_E, d_E	biosynthesis and degradation of endogenous miRNAs
c_X, d_X	biosynthesis and degradation of exogenous miRNAs
b_E, u_E	association and dissociation rate of endogenous-miRNA loaded RISC complexes
b_X, u_X	association and dissociation rate of exogenous-miRNA loaded RISC complexes
c, d_M	biosynthesis and degradation of the target mRNA
l	translation rate of the free target mRNA
l_A	translation rate of the RISC-bound mRNA
d_P	target protein decay rate
d_{cp}	deadenylation and decapping rate of the RISC-bound target mRNA
u_{Mp}	degradation of the deadenylated, decapped target mRNA and RISC recycling rate

Table 3: The 15 model reactions rates to be estimated. In addition, we may also need to estimate the total number of RISCs R and the inducible promoter leakage rate, resulting in a total of 17 parameters.

d_M to u_{Mp} while the exogenous-miRNA loaded RISC is recycled to A_X.

The reaction rates will be estimated from experimental measurements by a strategy which will be described in section 6.4.4. Given these reaction rates, we define the ODE system driving these dynamics as:

$$\begin{cases} \frac{dE}{dt} &= c_E - d_E E - b_E AE + u_E A_E \\ \frac{dX}{dt} &= c_X - d_X X - b_X AX + u_X A_X \\ \frac{dA_E}{dt} &= b_E AE - u_E A_E \\ \frac{dA_X}{dt} &= b_X AX - u_X A_X - b_{AM} A_X M + u_{M_P} M_P + d_M M_X \\ \frac{dM_X}{dt} &= b_{AM} A_X M - d_M M_X - d_{cp} M_X \\ \frac{dM}{dt} &= c - d_M M - b_{AM} A_X M \\ \frac{dP}{dt} &= lM + l_A M_X - d_P P \\ \frac{dM_P}{dt} &= d_{cp} M_X - u_{M_P} M_P \\ A &= R - (A_E + A_X + M_X + M_P) \end{cases}$$

6.4.3 *Steady state and initial conditions*

At the initial condition $t = 0$, the system is assumed to be at steady-state. There is no miRNA X at $t = 0$ and $c_X = 0, \forall t < 0$. However, this assumption is not very realistic because we know the inducible miRNA expression construct used here to be "leaky": the transcription rate of the exogenous miRNA before we induce its expression may take any value from 0% up to $\alpha \simeq 30\%$ of the rate when fully induced. In other words, if c_X^{max} is the transcription rate when the miRNA expression construct is fully induced,

$$\begin{cases} c_X \in [0, \alpha c_X^{max}], & t < 0 \\ c_X = c_X^{max}, & t \geqslant 0 \end{cases}$$

So, for the initial conditions, if we assume that the miRNA transcription is not leaky ($\forall t < 0, c_X = 0$), the initial conditions are:

$$\begin{cases} E(0) &= \frac{c_E}{d_E} \\ X(0) &= 0 \\ A_E(0) &= \frac{R}{\frac{d_E u_E}{b_E c_E} + 1} \\ A_X(0) &= 0 \\ M_X(0) &= 0 \\ M(0) &= \frac{c}{d_M} \\ P(0) &= \frac{lc}{d_M d_P} \\ M_P(0) &= 0 \end{cases}$$

If we assume, however that the construct is very leaky ($\forall t < 0, c_X = \alpha c_X^{max}$), the initial conditions are:

$$\begin{cases} E(0) &= \frac{c_E}{d_E} \\ X(0) &= \frac{\alpha c_X^{max}}{d_X} \\ A_E(0) &= \frac{k_E}{k_X} A_X(0) \\ A_X(0) &= \max\left(-k_X \frac{b+\sqrt{\Delta}}{2(k_E+k_X+1)}, -k_X \frac{b-\sqrt{\Delta}}{2(k_E+k_X+1)}\right) \\ M_X(0) &= \frac{c}{(d_M+d_{cp})\left(\frac{d_M}{b_{AM}A_X(0)}+1\right)} \\ M(0) &= \frac{c}{d_M+b_{AM}A_X(0)} \\ P(0) &= \frac{c}{(d_M+b_{AM}A_X(0))d_P}\left(1+\frac{l_A b_{AM}}{d_M+d_{cp}}A_X(0)\right) \\ M_P(0) &= \frac{c}{u_{M_P}\left(\frac{d_M}{d_{cp}}+1\right)\left(\frac{d_M}{b_{AM}A_X(0)}+1\right)} \end{cases}$$

where

$$\begin{aligned} k_E &= \frac{b_E c_E}{d_E u_E} \\ k_X &= \frac{b_X c_X}{d_X u_X} = \frac{\alpha b_X c_X^{max}}{d_X u_X} \\ b &= \left(\frac{k_E+1}{k_X}+1\right)\frac{d_M}{b_{AM}} + \frac{c}{d_M+d_{cp}}\left(\frac{d_{cp}}{u_{M_P}}+1\right) - R \\ \Delta &= b^2 + 4\left(\frac{k_E+1}{k_X}+1\right)\frac{d_M R}{b_{AM}} \end{aligned}$$

If we are unable to determine the leakage rate experimentally, we could assume that the initial conditions are uniformly distributed in the 9-dimensional cube whose boundaries are specified by the two sets of initial conditions above.

6.4.4 *Parameter estimation*

PRIOR ON THE PARAMETERS We determined plausible ranges for the model parameters from various biological constants measured in the literature. These ranges assume the exogenous miRNA X to be absent and are designed to exclude implausible parameter values rather than guessing the precise parameter values. On a time-unit of one hour (1h):

$$\begin{cases} b_E, b_X \in [10^{-4}, 10^3] \\ u_E, u_X \simeq .1 \in [10^{-3}, 10] \\ b_{AM} \in [10^{-4}, 10^3] \\ u_{M_P} \in [10^{-3}, 10^3], > d_M \\ d_{cp} \in [10^{-4}, 10^3] \\ R \in [10^2, 10^8] \\ c_E, c_X \in [10^2, 10^5] \\ c \in [10^{-2}, 2000] \\ d_E, d_X, d_M \simeq .07 \in [10^{-3}, 5], \\ l \in [1, 10^4] \\ l_A \in [0, l] \\ d_P \in [.1, .5] \end{cases}$$

OBSERVABLES AND A PRIORI MEASUREMENT ERROR Not all state variables of the detailed model illustrated on Figure 24 are easily mea-

surable. In the present study, we will obtain measurements of the following variables:

- $M + M_X + M_P$
- $X + A_X + M_X + M_P$
- P

The experimental techniques to measure the quantities above are designed to measure relative amounts, not absolute ones. It is therefore common practice to only measure changes in the quantities above — *i.e.* $\frac{Y(t)}{Y(0)}$ — which circumvents the difficulty of calibrating the measurements to obtain absolute concentrations.

$$\begin{cases} M + M_X + M_P & \pm 20\% \\ X + A_X + M_X + M_P & \pm 20\% \\ P & \pm 40\% \end{cases}$$

The relative amounts of target mRNA $M + M_X + M_P$ will be measured by qPCR, while the relative amounts of exogenous miRNAs $X + A_X + M_X + M_P$ will be measured using a qPCR protocol specially designed to quantify short RNAs [194, 35, 177, 188]. In practice, it should be achievable to perform 5 – 20 time-points and 3 – 10 replicates. Based on the system specifications of this section, our initial plan is to:

1. estimate the transcription rate c, translation rate l, mRNA decay rate d_M and protein decay rate d_P from transcription and translation inhibition experiments
2. obtain Bayesian estimates of the remaining 11 reaction rates from time-series measurements of the exogenous miRNA and the target gene mRNA and protein levels, using a sequential Monte Carlo approach [29], which would produce confidence intervals on the model parameters in addition to point estimates.

6.5 OUR DETAILED MODEL OF MIRNA REGULATION IS CONSISTENT WITH INDEPENDENT EXPERIMENTAL MEASUREMENTS

6.5.1 *Analytical analysis*

In the absence of the measurements we envision, we can still perform qualitative and steady-state analysis of the ODE system. This system is non-linear and admits a unique steady-state. Linearizing the system around that steady-state shows that it is stable given biologically meaningful parameters (positive rates and species concentrations).

Under the steady-state assumptions, one can examine under what conditions over-expressing a miRNA leads to down-regulation of the target mRNA. One can show that:

$$\frac{M'}{M} < 1 \Leftrightarrow d_M < u_{M_P}$$

where M' and M are the target mRNA levels at steady-state, in the presence and absence of the exogenous miRNA. Under the model's assumptions, and at steady-state, miRNAs decrease the levels of target

mRNAs if and only if the decay rate of deadenylated, decapped mRNAs is larger compared to capped, polyadenylated mRNAs. This is in agreement with the biology of mRNA turn-over as the 5′ cap and the polyA tail protect mRNAs from digestion by exonucleases. Interestingly, under the model's assumptions, the converse is also true which can lead to interesting speculations: if RISC binding to the mRNA could somehow relocate the mRNA to a compartment where it is not accessible to exonucleases, it is expected that miRNAs could up-regulate their targets. A miRNA-mediated up-regulation of the target mRNA was previously reported by Vasudevan et al. [220], and it may be interesting to think about the underlying mechanism with this idea in mind.

Similarly, one could ask under what conditions changes in protein levels are expected to be larger than changes in mRNA levels. At steady-state, one can show that:

$$\frac{P'}{P}\frac{M}{M'} < 1 \Leftrightarrow \frac{l_A}{l}\frac{u_{Mp}}{u_{Mp}+d_{cp}} < 1$$

where P' and P are the protein levels in the presence and absence of the exogenous miRNA. Because we have no reason to expect that the translation rate l_A of RISC bound mRNAs is larger than the translation of RISC free mRNAs l, we have $l_A \leqslant l$. Recent work by Fabian et al. [64] has proposed that the RISC interacts with the PolyA Binding Protein (PABP), thereby interacting with the circularization of the mRNA, which should have a negative effect on the translation rate, so we can reasonably assume that $l_A < l$. The term on the right must be smaller than one since all rates are positive. Therefore, under the assumptions of the model, target proteins should experience stronger down-regulation than target mRNAs, which is in agreement with the experimental literature (see Zipprich et al. [240] for instance). According to the model, this would also be the case if miRNAs did not lead to deadenylation and decapping of their target mRNAs ($d_{cp} = 0$), or if miRNAs caused an instant disintegration of the target mRNAs ($u_{Mp} \to \infty$).

In conclusion, independent of the precise values of the reactions, our model can qualitatively reproduce fundamental properties of miRNA-mediated gene regulation.

6.5.2 *Estimating the rates of exogenous siRNA-Ago complex formation from Fluorescence Cross-Correlation Spectroscopy measurements*

Using Fluorescence Cross-Correlation Spectroscopy (FCCS), Ohrt et al. [163] performed several time-series measurements of the fraction of Ago-bound siRNA following the microinjection of labeled siRNA. In addition, using the same technique, the fraction of Ago bound to the microinjected siRNA was also quantified over time. We sought to analyze this dataset with the goal of obtaining coarse-grained estimates of the RISC loading and unloading rates as well as of the decay rate of small RNAs to be used as a starting point to estimate the remaining 11 parameters of our model.

To do so, we devised a simplified model of small RNA loading into RISC illustrated on Figure 25. An amount of siRNA X_0 is microinjected at time $t = 0$. The free siRNA X degrades at the rate $d_X X$ and loads

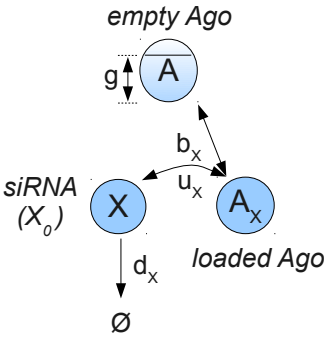


Figure 25: A simple model of a microinjected siRNA associating and dissociating with Ago.

into a free Ago A with rate b_X to form a complex A_X that dissociates at rate u_X.

We assumed that only a constant fraction g of all Ago A is free to associate with a free siRNA X while the remaining fraction $1-g$ of Agos A is bound to endogenous miRNAs. This is a simplification from the detailed model introduced in section 6.4.2 where the competition between endogenous and exogenous miRNAs would result in g implicitly changing over time as exogenous miRNAs make room for themselves to load into Ago.

In addition, this simplified model does not take ternary complex formation into account. Therefore, from the different FCCS time-series performed by Ohrt et al. [163], we focused on the cytoplasmic measurement as miRNA regulation is thought to be happening in the cytoplasm, following of the microinjection of the siTK3 siRNA. We chose to focus on this siRNA since it was designed to perfectly target the Renilla luciferase mRNA used in these experiments: compared to siRNA with imperfect complementarity, we expect this siRNA to trigger rapid degradation of the target mRNA by cleavage as opposed to binding the target for a longer time and trigger the slower miRNA pathway (GW182, deadenylation, decapping, etc.). In other words, with the siTK3 siRNA, the time spent by the siRNA in a ternary Ago-siRNA-mRNA complex should be minimal, which is more in agreement with the simple model we use here.

The model sketched on Figure 25 translates into the following ODE system:

$$\begin{cases} \frac{dX}{dt} &= -d_X X - b_X g(R - A_X) + u_X A_X \\ \frac{dA_X}{dt} &= b_X g(R - A_X)X - u_X A_X \end{cases} \tag{6.5}$$

where R is the total amount of RISC and the initial conditions are $X(0) = X_0$, $A_X(0) = 0$. However, Ohrt et al. [163] measured the fraction of Ago in complex $\rho = \frac{A_X}{R}$ and the fraction of siRNA in complex

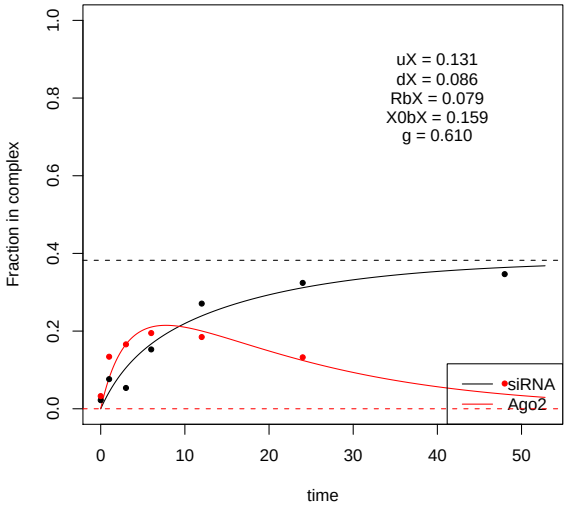


Figure 26: Fitting the cytoplasmic siTK3 Fluorescence Cross-Correlation Spectroscopy time-series of Ohrt et al. [163]. Red and black dots represent the measured fraction of Ago2 and siRNA molecules in complex over time (in hours), while the full red and black curves show the fitted time-course of the model defined by equation 6.6. Finally, the dotted red and black lines represent the predicted siRNA and Ago2 steady-state of the model while the best-fitted parameter values are printed in the top right corner of the figure (hourly rates).

$\gamma = \frac{A_X}{A_X+X}$. Substituting ρ and γ in equation (6.5) results in a new ODE system:

$$\begin{cases} \frac{d\gamma}{dt} &= d_X\gamma(1-\gamma) + b_X R(g-\rho)(1-\gamma) - u_X\gamma \\ \frac{d\rho}{dt} &= b_X R(g-\rho)\rho\frac{1-\gamma}{\gamma} - u_X\rho \end{cases} \tag{6.6}$$

Under the initial conditions $(\gamma(0) = 0, \rho(0) = 0)$, $\left.\frac{d\rho}{dt}\right|_{t=0}$ is undefined. But going back to the definition of $\rho = \frac{A_X}{R}$, we get:

$$\left.\frac{d\rho}{dt}\right|_{t=0} = \frac{1}{R}\left.\frac{dA_X}{dt}\right|_{t=0} = b_X g X_0 \tag{6.7}$$

The model defined by equations (6.6) and (6.7) has 5 parameters: $g, u_X, d_X, b_X R$ and $X_0 b_X$. We obtained estimates of these parameters by minimizing the squared model prediction error using the method of Nelder and Mead [160]. The prediction error was minimized starting from 500 random parameters sets. Figure 26 shows the FCCS time-series data together with the prediction of the model based on the best-fitted parameters. Note that the fraction of siRNA in complex does not converge to 0 but .4, indicating that the total amount of siRNA decreases about $\frac{1}{.4} = 2.5$ times as fast as the amount of Ago-bound siRNA, consistent with Ago-loaded siRNAs being protected from degradation.

The model is able to accurately fit the measurements and the values of the best-fitted parameters were consistent across runs, except

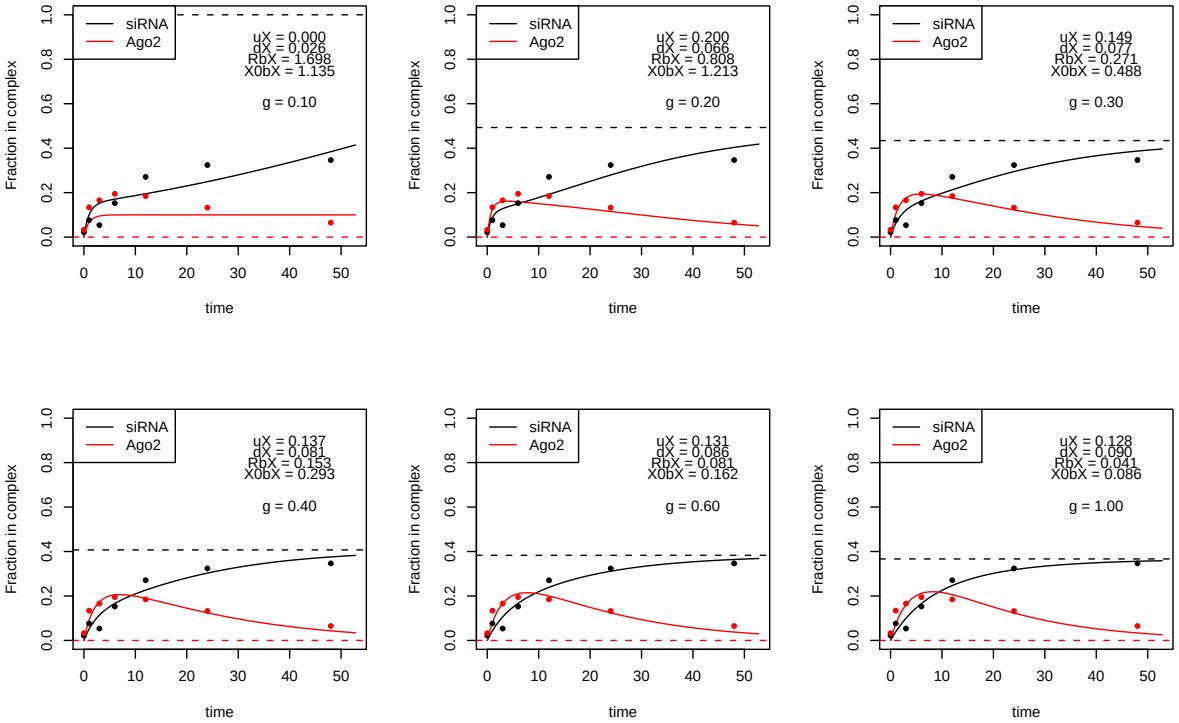


Figure 27: Best fitting the model parameters while fixing g to .1, .2, .3, .4, .6 or 1.

for the fraction g of available Ago which varied significantly from one run to next. This is best shown by Figure 27, where the model parameters were best-fitted starting with 500 random parameter values while constraining g to a value between 0 and 1. While $g = .61$ fitted the measurement best, any value between .3 and 1 resulted in models that fitted the data just as well. This is in contrast with the 4 remaining parameters, which took comparable values independent of g changing between .3 and 1, suggesting that the data is not very informative regarding the fraction of Ago that is available to the microinjected siRNA. As long as g is larger than the maximum fraction of Ago in complex — 20%, which occurs ~8h after transfection here — parameters can be found to fit the FCCS measurements.

As for the remaining parameters, the siRNA decay rate d_X is estimated to .077 – .09 per hour, which corresponds to a half-life of ~8h. This is comparable to the half-lives of most mRNAs in B-cells which were measured to be in the 3h – 11h range [68]. The siRNA-Ago dissociation rate was estimated to be around $\sim .13h^{-1}$ per RISC complex. Finally, under the model's assumptions, there was an estimated average of $\frac{X_0}{R} \simeq 2$ microinjected siRNAs per Ago. However, these estimates need to be taken with a grain of salt, as there are likely to be biased. For instance, we do not model Ago-siRNA-mRNA ternary complex formation, which probably increases the amount of time siRNAs spend in Ago. Therefore, u_X is probably higher than we estimated it here. Nevertheless, the parameter estimates we obtained fit reasonably with our expectations and make a good starting point to estimate the remaining parameters of the model.

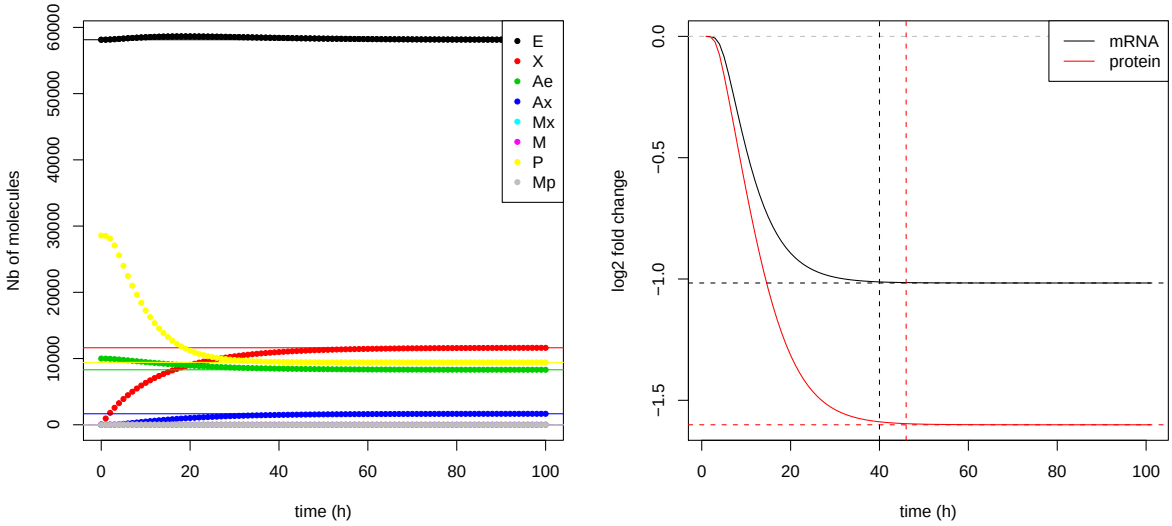


Figure 28: Simulating the induction of an exogenous miRNA X at time 0. Left: Simulated evolution of the number of molecules per cell (y-axis) over time (x-axis, in hours). Full horizontal lines represent the steady-state values of the different model species in the presence of the exogenous miRNA. Right: Simulated log2 fold changes (y-axis) in the protein (red) and mRNA (black) levels of the target gene with respect to time after miRNA induction (x-axis, in hours). The horizontal dotted lines represent the log2 fold changes of the protein and mRNA in the presence of the exogenous miRNA steady-state, while the vertical dotted lines mark the time points at which the steady-states are reached.

6.5.3 *Simulations and timing*

Prior to performing the time-series measurements described in section 6.4.4, we set to study the behavior of the model using guessed but plausible parameters. To do so, for each of the 15 parameter, we plugged in the middle of the confidence intervals defined in section 6.4.4 into the model, except for u_X and d_X, for which we used the values determined in section 6.5.2.

Using these estimates, we simulated time-courses for all the model species following the induction of the exogenous miRNA at time 0. The simulation results are shown on left panel of Figure 28. In these simulations, we observe that the induction of the exogenous miRNA causes the amount of free miRNA X and Ago-bound miRNA A_X to grow over time until it reaches a new steady-state ~40h after induction, resulting in an expected drop of the levels of the target protein P. Interestingly, with a ratio of ~6 free endogenous miRNAs per RISC-loaded endogenous miRNA, the majority of miRNAs are *not* loaded in a RISC in these simulations. This is also observed in the FCCS measurements of Ohrt et al. [163], where the fraction of loaded siRNA never exceeds 50%. Finally, in spite of the target protein levels dropping to a third of their original levels, the amount of free and bound miRNAs appears to be barely affected, which may indicate that it could be possible for a miRNA to repress a target gene significantly, without having to take over a dramatic proportion of RISCs.

Parameter	Effect on protein regulation	Effect on mRNA regulation
c	slower	slower
d_M	weaker, faster	weaker, faster
l	stronger and faster, none if $l_A \alpha l$	none
d_P	faster	none
b_{AM}	stronger, faster	stronger, faster
d_{cp}	stronger	stronger, faster
u_{Mp}	none	stronger, faster
l_A	weaker, faster	none

Table 4: Perturbation analysis of gene-dependent parameters. The effect of turning up each parameter on the regulation of the protein and mRNA levels of the target gene are indicated in their respective columns.

In addition, we simulated the log2 fold changes in the protein and mRNA of the target gene, shown on the right panel of Figure 28. There, the protein and mRNA levels of the target gene drop together and reach the new steady-state almost simultaneously, ~40h after miRNA induction. That the protein and mRNA reach the steady-state at the same time is mostly due to the fact that we are simulating a protein with a half-life of 3h ($d_P = .25$): with a protein with a half-life of 24h, the steady-state is reached after 100h (not shown). On the mRNA level, however, this time-scale of 40h is in agreement with the typical time-scale of miRNA transfection experiments (see Figure 4 of Khan et al. [118] for instance). In addition, we see that protein levels change more compared to mRNA levels, with protein and mRNA levels dropping to 33% and 50% of their original levels, respectively. This is expected from the results of section 6.5.1, and again in agreement with the established knowledge in miRNA biology [66, 62].

6.5.4 *Perturbation analysis*

In order to get a better understanding of the effect the different gene-dependent parameters may have on the time-scale and magnitude of miRNA-mediated gene regulation, we performed a parameter perturbation analysis. We individually tuned each parameter up and down while keeping all other parameters constant, which allowed us to explore the regulatory consequences of changing the parameters. The results of this analysis are shown on Figure 29. For instance, the panel in the second column of the second row looks at the effect of the miRNA-loaded RISC – target mRNA binding rate b_{AM}. As we increase b_{AM} from 10^{-5} to 5.10^{-4}, the magnitude of the down-regulation increases both at the protein and mRNA levels, while the steady is reached sooner. In other words, the better the miRNA affinity to its target, the faster and the stronger the regulation.

Table 4 summarizes the perturbation analysis for all parameters. It appears clearly that all gene-dependent parameters influence how fast target genes are regulated at the protein level, at the mRNA level, or both, as well as the strength of the regulation. For instance, our model predicts that mRNAs harboring several high-affinity miRNA binding sites — which should result in a high b_{AM} — are stronger

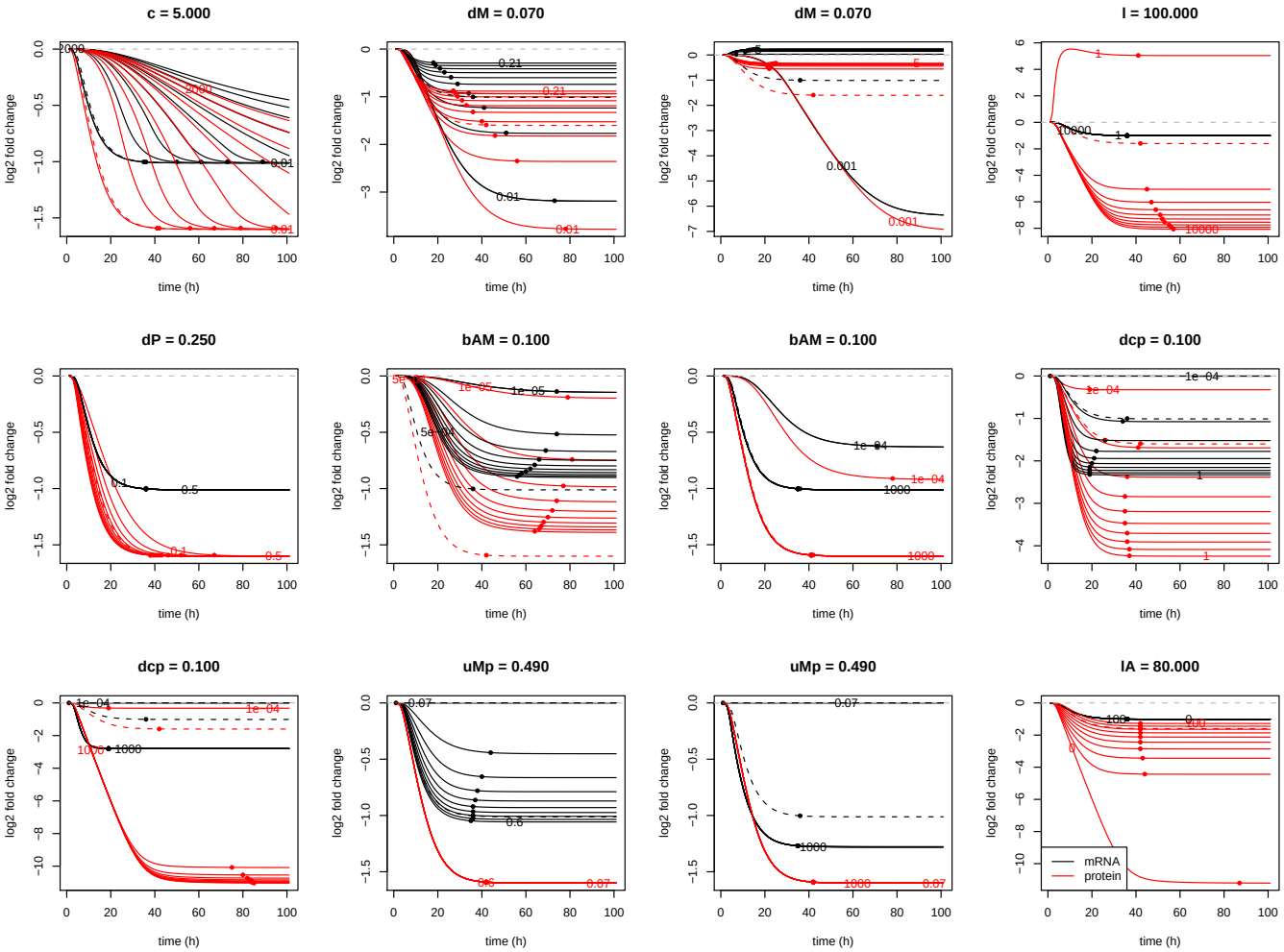


Figure 29: Parameter perturbation analysis of the detailed ODE model introduced in section 6.4.2. Each panel is concerned with the effect of an individual parameter — specified in the title together with its original value — on the kinetics of miRNA-mediated gene regulation. The dotted red and black lines represent the log2 fold change in the protein and mRNA levels of the target gene following the induction of the miRNA under the original parameters. These parameters are then progressively turned up and down in constant steps around two extreme values which appear on the figures next to the corresponding trajectories. The black and red dots mark the time points at which protein and mRNA reach steady-state levels.

regulated. This is consistent with observations from miRNA transfection microarray experiments, where the extent of the pairing to the 5′ end of the miRNA and the number of miRNA binding sites correlates with the amount of mRNA degradation [83]. In addition, our model also predicts that slow-decaying mRNAs (small d_M) should experience a stronger regulation than fast-decaying mRNAs, which remains to be tested in high-throughput.

When it comes to the time-scale of miRNA-induced regulation, for most parameters, tuning the parameter in such way that the regulatory effect increases also makes the regulation faster. For instance, the model predicts that fast-decaying proteins (high d_P) translated from slowly transcribed mRNAs (small c) carrying several high-affinity miRNA binding sites (high b_{AM}) should experience faster down-regulation. Again, these predictions remain to be tested.

Exceptions to the `strong regulation = fast regulation` pattern are the mRNA decay rate d_M and the translation rate of Ago2-bound mRNAs l_A, which highlight a limitation of this kind of "single parameter" analysis. The analysis performed in this section does not take dependencies between parameters into account, which leads to problems given the parametrization we used here. For instance, if we turn up the mRNA decay d_M enough, it will at some point reach the value $d_M = u_{Mp}$. In that case, the model predicts that miRNAs have no effect on mRNA levels (see section 6.5.1), which highlights the source of the problem: the degradation rate u_{Mp} of deadenylated, decapped mRNAs is probably not a constant, but varies relatively to d_M. If a given polyadenylated, capped mRNA degrades fast, then we expect that it will degrade even faster when deadenylated and decapped. Therefore, a more sound parametrization of the model may be $u_{Mp} = k_{cp} d_M$ or $u_{Mp} = k_{cp} + d_M$, where k_{cp} represents the increase in mRNA decay resulting from loosing the cap structure and the polyA-tail. In section 6.7.3, we will briefly introduce a method that may be useful at finding all parameter dependencies of this type so they can be eliminated, which should result in a better model.

6.6 CONCLUSION

In this chapter, we analyzed high-throughput quantitative proteomics and transcriptomics measurements following miRNA transfection. Using competing, simple models of the mechanism of miRNA-mediated gene regulation, we showed that changes at protein levels measured by state of the art SILAC and pSILAC miRNA target identification protocols are far from steady-state and do not accurately reflect the ultimate effect of miRNA regulation. This may explain why functional miRNA binding sites do not necessarily lead to down-regulated protein levels in a miRNA transfection experiment, and suggests that the precise kinetics of miRNA-mediated regulation need to be taken into account when designing and analyzing such experiments.

For this purpose, we introduced a detailed ODE model of the kinetics of miRNA-mediated gene regulation and sketched a strategy to estimate the model parameters from experimental measurements. We presented different open problems that this model may help resolving and showed that the steady-state predictions of the model are consistent with key observations of the miRNA literature. As a preliminary to estimating all the parameters of the model, we set to estimate pa-

rameters related to the RISC loading and unloading of small RNAs from published measurements. Plugging plausible parameters values into the detailed ODE model and simulating a miRNA induction experiment, we obtained time-courses where the time scale and the magnitude of miRNA-mediated gene regulation is consistent with those of typical miRNA transfection experiments. Finally, all gene-dependent parameters appear to inclluence either the timing or the magnitude, or both aspects of miRNA regulation.

6.7 FUTURE WORK

A weakness in those latest simulations is that parameter estimates were at best estimated from measurements performed in a different context using coarse-grained models, or at worse guessed from published measurements in the literature. As a preliminary to interpreting the model predictions further, the parameters need to be estimated more rigorously, which may require complementary experimental approaches and/or changes to the model in order to make these parameters identifiable under the inherent constraints and limitations of experimental approaches. In addition, complementary analyses are needed to confirm the reason that justifies building such a model, to check the validity of the model's assumptions, and to validate the model's predictions.

6.7.1 *Confirming that a model of the kinetics of miRNA-mediated gene regulation is necessary*

That the kinetics and transient of miRNA-mediated gene regulation need to be taken into account when designing and analyzing high-throughput quantitative proteomics experiments was so far only established on miRNA transfection datasets where a miRNA is transiently over-expressed before being eliminated by decay. For this reason, proteomics and transcriptomics measurements must be performed 2 to 3 days following miRNA transfection. A complementary approach when it comes to miRNA target identification consists in stably over-expressing or knocking out a miRNA by genetic means. This allows for much longer time intervals between the induction or knock-out of the miRNA and the proteomics and transcriptomics measurements. Therefore, we expect the proteomics and transcriptomics measurements from such systems to be at steady-state. Baek et al. [8] and Selbach et al. [191] performed miRNA genetic knock-out experiments and the corresponding proteomics and transcriptomics measurements could be used to confirm our hypothesis by examining the properties of miRNA binding sites leading to up-regulated protein levels, and by performing the model comparison analysis of section 6.3.4.

In addition, Guo et al. [88] recently published genome-wide ribosome profiles following miRNA transfection or knock-out. Such profiles provide a direct read out on miRNA-mediated translation regulation and it would therefore be interesting to examine whether miRNA binding sites leading to lower translation rates share the properties of functional miRNA binding sites.

6.7.2 *Checking model assumptions*

The present model relies on a set of explicit and implicit assumptions, which need to be validated experimentally prior to trusting the model predictions. Such validation experiments include:

- Making sure Ago levels do not change upon inducing the miRNA (qPCR, western blot)
- Making sure that the production and degradation rates of endogenous miRNAs do not change when inducing the new miRNA
- After estimating the rates c_M, d_M, l, d_P of the target gene in the absence of the 3'UTR recognized by the miRNA, induces the expression of exogenous miRNA and making sure these parameters stay constant over time: they could change because the induced miRNA may interfere with the transcription, mRNA degradation, translation or protein turn-over machinery, which would flaw parameter estimation.

In addition, one may need to add a nuclear mRNA compartment to the model or remove cell nuclei experimentally prior to quantifying miRNAs, as nuclear-located miRNAs cannot interfere with translation regulation.

6.7.3 *Parameter estimation*

IDENTIFIABILITY So far, we have not looked at the question of theoretical identifiability: given error-free, high-frequency sampled time-course measurements of the observables of the system, are the parameters uniquely identifiable in principle?

Let $\{x_i, i = 1, \dots, n\}$ be the error-free measurements performed on time-points $\{t_i, i = 1, \dots, n\}$, and let $\{x(t_i, \theta), i = 1, \dots, n\}$ be the corresponding model predictions under the parameters θ. We can infer the model parameters θ^* that are most in agreement with the measurements by minimizing the squared model prediction error:

$$\theta^* = \text{argmin}_\theta \sum_{i=1}^{n} (x(t_i, \theta) - x_i)^2$$

But assuming we have obtained such parameters θ^* and reached the global minimum of the squared model prediction error landscape, how can we make sure that no other parameter combinations would just as accurately reproduce the data? In other words, is it possible to find a perturbation to the parameters that would result in different parameters but with an identical squared model prediction error?

Figure 30 shows the model prediction error landscape for the toy error function $\epsilon(x, y) = (xy - 1)^2$. This error function has two parameters x and y, while 1 represents the measurements. For all (x, y) such as $x = \frac{1}{y}$, we have $\epsilon(x, y) = 0$. Therefore, this model prediction error function does not provide information on the individual values of x and y, only on the value of the product xy. This mirrors the observation of Table 4 in section 6.5.4 where tuning l_A in proportion to l had no effect on the time-course of protein and mRNAs following miRNA induction. However, what if the model prediction error does not have a closed-form and cannot be studied analytically? This is the case in our

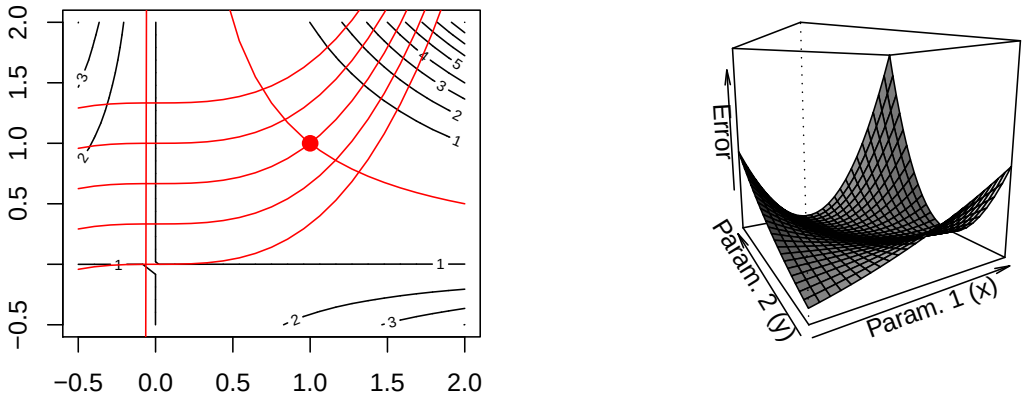


Figure 30: A toy example of a model prediction error landscape. Shown are contour (left) and perspective (right) representations of function $\epsilon(x,y) = (xy-1)^2$, where $x > 0$ and $y > 0$ are the "model parameters", 1 is the "data" and $\epsilon(x,y)$ is the model prediction error.

detailed model, where the model prediction eror $\sum_{i=1}^{n}(x(t_i,\theta)-x_i)^2$ needs to be evaluated numerically. In this case, we can still attempt to find local dependencies between parameters using a numerical approach.

Let $\epsilon(\theta) = \sum_{i=1}^{n}(x(t_i,\theta)-x_i)^2$ be the squared model prediction error. $\frac{d\epsilon}{d\theta}\big|_{\theta=\theta^*} = 0$ since θ^* is an extremum of ϵ. Therefore, the Hessian of ϵ computed at θ^*

$$\left.\frac{d^2\epsilon}{d\theta^2}\right|_{\theta=\theta^*}$$

summarizes the information regarding how small perturbations to the parameters around θ^* affect the squared model prediction error. Finally, an eigenvector decomposition of the Hessian matrix reveals along the directions along which perturbations to the parameters have strongest and weakest effects on the squared model prediction error. In the case of $\epsilon(x,y) = (xy-1)^2$ and $\theta^* = (1,1)$, we have

$$\left.\frac{d^2\epsilon}{d\theta^2}\right|_{\theta=\theta^*} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

The eigenvalues of this Hessian are 2 and 0, corresponding to the vectors $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$ respectively. In other words, perturbing x and y in the same direction by the same small amount will result in a maximum increase of the model square prediction error while perturbing x and y in opposite directions by the same small amount will leave the squared model prediction error unchanged, as shown by simple observation of the definition of ϵ above. Therefore, the eigenvectors with corresponding egenvalues close to 0 reveal local dependencies between parameters.

Preliminary application of this local perturbation analysis around the parameters introduced in section 6.4.4 shows that the parameters of the model are far from being identifiable from time-courses of just three observables. But upon taking the rates governing RISC loading and unloading of miRNAs out of the analysis, parameters best-fitting the experimental measurements become uniquely determined, except for the translation rate of the Ago-bound mRNA (l_A) and the decay rate of endogenous miRNAs (d_E) (not shown). This indicates that parameters would be close to being identifiable if the biophysics of miRNA-RISC complex formation were known.

In the future, we plan to refine this analysis along the ideas exposed in Brown and Sethna [26] with two goals in mind. First of all, we will seek to determine individual, non-identifiable parameters that only have a negligible influence on the time-course of the model species we are interested in. For instance, the rates governing endogenous miRNA-RISC complex formation are likely to be hard to identify while having only a small influence on the time-courses of mRNA and protein levels. But we have limited interest in these rates beyond modeling the competition of small RNAs for RISCs, so we may not need precise estimates of the corresponding individual parameters anyway. Second, we will determine collections of parameters that are non-identifiable under the current observables so that we can come up with alternative solutions to estimate them. Possible solutions to estimating such dependent parameters are described in the rest of the present subsection.

SIMPLIFY THE SYSTEM MAKE IT IDENTIFIABLE Parameters occurring in eigenvectors corresponding to small eigenvalues are likely not to be identifiable given the observables. One solution may be to simplify the model by collapsing the corresponding reactions into fewer, abstracter reactions.

Pushing the logic of the argument to the extreme, we need to check how much difference it would make to follow Khanin and Higham [119] in assuming that miRNA induce simultaneous changes in the translation and mRNA decay rates, as opposed to explicitly modeling the competition of small RNAs for RISC, the initial translation repression and the subsequent delayed mRNA degradation.

Finally, we need to examine whether time-scale separation can be used to eliminate certain states and rates. For instance, can we assume that Ago-miRNA-association/dissociation is fast and is therefore at equilibrium compared to other reactions? Maybe a careful comparison between the time-courses of Ago loading from Ohrt et al. [163] and mRNA down-regulation of Wang and Wang [226] can help answering this question. More generally, it may be important to screen all the reactions in our detailed model to check if it can be simplified further.

MEASURE ADDITIONAL QUANTITIES Our preliminary local parameter perturbation analysis (section 6.7.3) suggested that if we knew the biophysics of miRNA-RISC complex formation, we may be very close to being able to identify the remaining parameters. This suggests that adding biophysics measurements — the fraction of bound Ago-miRNA over time for instance —- to the observables may be crucial in solving the parameter estimation problem. Or maybe it would be possible to measure subpopulations of mRNAs, such as the fraction of deadenylated target mRNA among all reporter mRNAs.

APPLYING OTHER PERTURBATIONS The exogenous miRNA induction rate c_X is easily tunable experimentally. Therefore, we will investigate whether differential inductions of the exogenous miRNA — *i.e.* performing several time-course measurements of the observables under different c_X but keeping all other rates constant — can help identifying the parameters. It may also be possible to knock down one or several component of the system, or induce the expression of the exogenous miRNA for a limited period of time only. We will determine which of these perturbations are most informative using the framework developed by Bandara et al. [10].

PERFORMING SINGLE-CELL MEASUREMENTS The stochastic fluctuations of the molecules cross-reacting in a genetic network carry information on the reaction rates of the underlying network, including on rates that may be difficult to measure by other means. We could profit from this effect by attempting to perform parameter estimation within a framework that is able to extract information from cell-to-cell fluctuations in the number of molecules [158], as opposed the option we have been considering so far that only uses information from cell population average measurements.

Experimentally, this would involve the combination of FACS (Fluorescence-Activated Cell Sorting) measurements of luciferase intensities, and single molecule RNA FISH measurements of the target mRNA [174]. FACS is commonly available in molecular biology institutions while single molecular RNA FISH only requires standard wide-field microscopy. Specific experimental procedures were developed for single-cell measurements of miRNAs [33].

6.7.4 *Validating the model*

Future work will also put the model prediction in the context of the regulatory consequences of knock-downs of RNAi components on target gene mRNA and protein levels in Drosophila, such as those performed by Eulalio et al. [60]. Such experimental results would allow us to check whether the model generalizes qualitatively to biological contexts that were not used to fit the parameters and thereby test if the model is also predictive in biological conditions it has never "experienced".

Finally, to quantitatively validate the model, one may test the predictions of the parameter perturbation analysis in section 6.5.4 on high-throughput datasets. For instance, the mRNA decay rates were measured by several studies Yang et al. [234], Cheadle et al. [34], Friedel et al. [68], Iwamoto et al. [104] as well as human protein decay rates [48]. Using such measurements, one could test whether the predicted effect of the mRNA decay rates on the time-scale of miRNA regulation can be observed in large microarray time-series of miRNA overexpression [226, 118]. One could also test whether the predicted effect of the mRNA and protein decay rates on the magnitude of miRNA regulation is in agreement with the measured changes in protein levels following miRNA transfection [191, 8]. Doing so would increase the confidence in the model, or highlight discrepancies indicative of an overlooked phenomenon in miRNA regulation. Either ways, the process would lead to a better understanding of miRNA biology.

SUPPLEMENTARY MATERIAL TO CHAPTER 2

A.1 SUPPLEMENTARY METHODS

A.1.1 *Bioinformatics analyses*

We imported the CEL files from the Affymetrix Mouse Genome 430 2.0 Array into the R software [173] by using the BioConductor affy package [74]. The probe set intensities were then background-corrected, adjusted for nonspecific binding, and quantile normalized with the GCRMA algorithm [233]. Probe sets with more than 2 probes mapping ambiguously (more than 1 match) to the genome were discarded, as were probe sets that mapped to multiple genes. We then collected all probe sets matching a given gene, and we selected for further analysis the Refseq transcript with median 3'UTR length corresponding to that gene. The log-intensities of probe sets mapping to the gene were averaged to obtain the expression level per Refseq transcript. Finally, we used Limma [199] to estimate the fold change and the P value of the difference in expression between 375KO and wild type for each transcript.

To investigate whether the transcripts responding to a particular treatment are enriched or depleted in matches to the miR-375 miRNA, we ranked the transcripts according to their estimated fold change. We labeled the 5% most down-regulated transcripts as "down" and the 5% most up-regulated transcripts as "up". What one means by "miRNA seed" varies to some extent from one study to another. Most commonly this term refers to positions 1-8, 1-7, 2-8, or 2-7 of the miRNA. Because the effect on mRNA stability depends on the extent of miRNA-mRNA sequence complementarity, we generally want to separately analyze putative sites with different degrees of complementarity. miR-375 has, however, a CG dinucleotide at positions 7-8, leading to a very low number of sites that are complementary to positions 2-8 or 1-8 of this miRNA. Because of the high variance associated with these low numbers of sites, we used for our analysis only sites that are complementary to positions 1-7 (not including those that are also complementary to positions 1-8) of miR-375. We call the 1-7 miR-375 seed complementary sequence with a mismatch at position 8 the "miR-375 motif". We counted how many times the miR-375 motif occurred in the 3'UTRs of transcripts labeled as "up" or "down". This number is represented in the plots as a black dot. To assess whether the number of miR-375 motifs is unusually high or low compared with what would be expected for a "random miRNA", we computed a "background" motif count distribution as follows. We considered all possible "random miRNA seeds." These are all of the possible octamers. For each of these, we determined the number of occurrences of "background motifs" in all of the 3'UTRs of transcripts monitored by the microarray. These are all 3'UTR positions that match perfectly the 1-7, but not the 8th position of the "random miRNA". We then selected the 5% of these background motifs (i.e., 3,277) whose number of occurrences in the entire set of 3'UTRs was closest to that of the miR-375 motif. We computed the

expected number of occurrences of these in the 3'UTRs of the "up" and "down" transcripts. This was defined as observed occurrences of the background motifs in the "up" or "down" transcripts 3'UTRs observed occurrences of the miR-375 motif in the entire 3'UTRs set/observed occurrences of the background motifs in the entire 3'UTRs set. The distribution of the number of occurrences of background motifs is represented as follows: the blue boxes show the interquartile range and the red line the median. The range bounded by the black whiskers indicates the interval that is 1.5 times the interquartile range, and the red dots show all background motifs whose number of occurrences does not fall within this range. The black dot represents the count of miR-375 motifs. The P value of the enrichment is given by the fraction of the background motifs that have at least as many occurrences as the miR-375 motif in the "up" transcripts, and the significance is represented by the location of the black dot within the box plot representing the distribution of background motifs. Finally, the P value of the depletion is given by the fraction of the background motifs that have at most as many occurrences as the miR-375 motif in the "down" transcripts.

A.1.2 *Isolated Islet Secretion and Capacitance Measurements*

Islet secretion studies were performed on size-matched islets isolated from 10-week-old animals following collagenase digestion and overnight culture and performed as described [170]. Exocytosis of secretory granules was monitored in single α or β cells by capacitance measurements as described previously [170, 56]. The measurements were performed in the standard whole-cell configuration of the patch-clamp technique at 32-33°C and the identity of β cells was subsequently confirmed after the experiment by immunocytochemistry.

A.2 SUPPLEMENTARY FIGURES

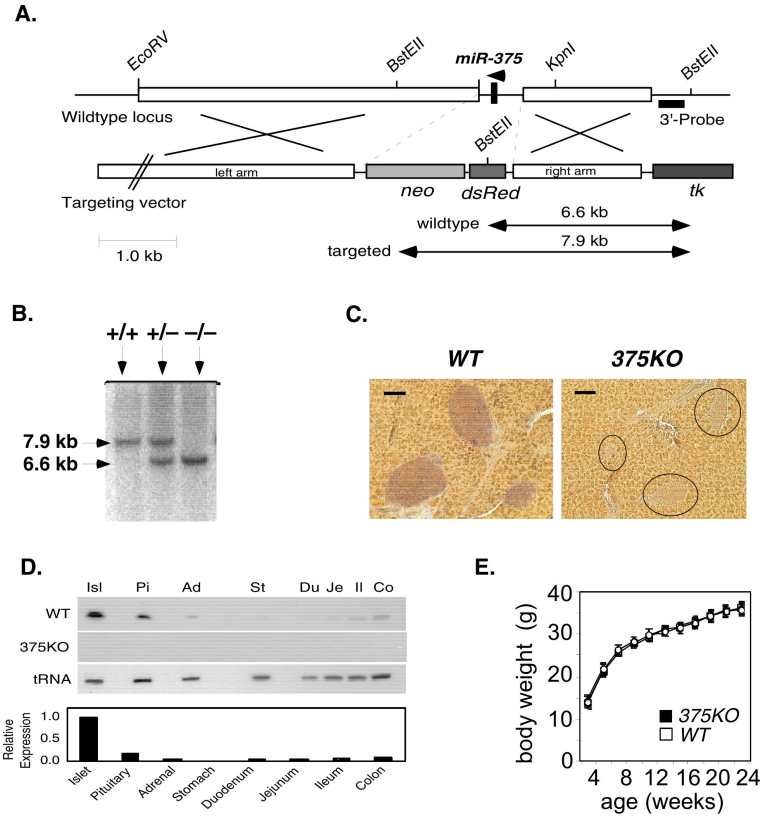


Figure 31: Deletion of the miR-375 gene by homologous recombination. (A) Targeting strategy for deletion of the miR-375 locus by replacement with dsRed cDNA and the neomycin (Neo)-resistance cassette by homologous recombination in ES cells. Targeting arms are shown as white boxes, and the probe 3′ to the right targeting arm that was used for Southern blot analysis is shown as a black bar. No fluorescence signal was observed from dsRed of heterozygous or null mice. (B) Analysis of genomic DNA from wild-type (+/+), miR-375 heterozygous (+/-), and miR-375 homozygous (-/-) mice after digestion with BstEII. (C) In situ hybridization in pancreatic sections from wild-type (WT) and mutant (375KO) mice with a probe for miR-375. Black circles indicate islets in miR-375-null mice. (Bar, 25 μm.) (D) Northern blot of total RNA isolated from 10-week-old WT and 375KO tissues: pancreatic islets (Isl), pituitary gland (Pi), adrenal gland (Ad), stomach (St), duodenum (Du), jejunum (Je), ileum (Il), and colon (Co). Blot was reprobed for tRNA as a loading control and quantified by densitometry. (E) Growth curve of 375KO and wild-type mice.

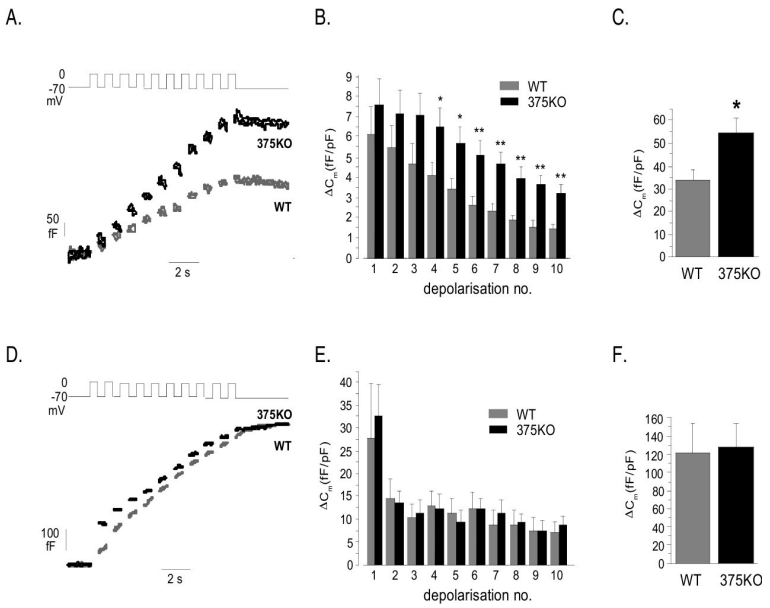


Figure 32: Single-cell capacitance measurements in pancreatic α and β cells of 375KO and littermate control mice. (A) Secretion was evoked by a train of depolarizations from -70 mV to 0 mV in isolated β cells from 10-week-old male 375KO (black signals) and wild-type (gray signals) mice. (B) Mean increase in membrane capacitance of isolated β cells elicited by individual depolarizations of the train ($\Delta C_{m,n} - \Delta C_{m,n-1}$) displayed against the pulse number (n). (C) Total mean increase in membrane capacitance elicited by individual depolarizations of pancreatic β cells from 375KO and wild-type mice. (D-F) As in A-C but using α cells. *, P < 0.01; **, P < 0.001.

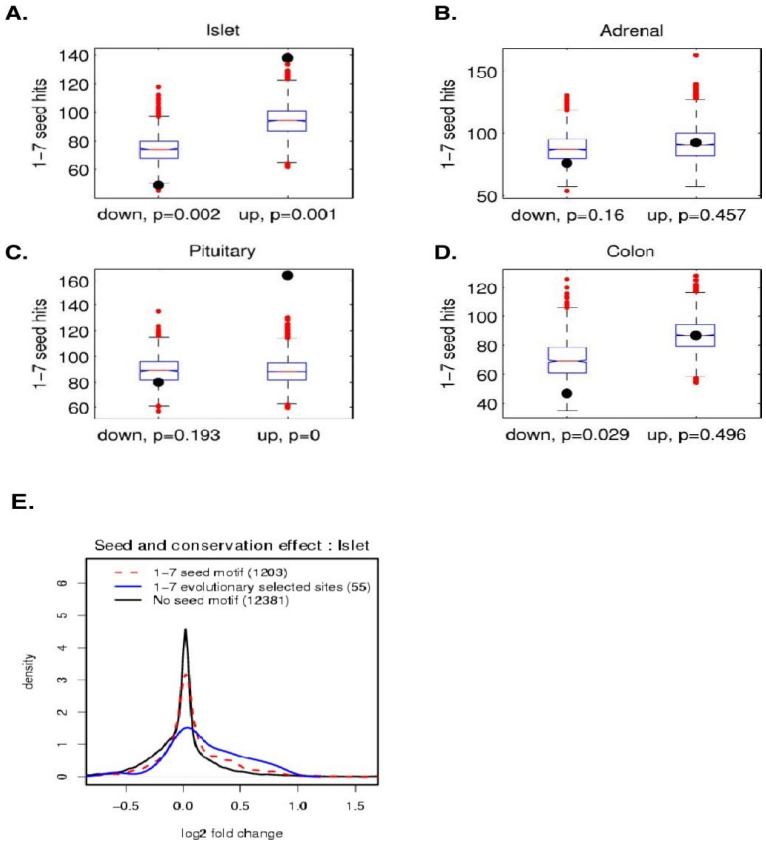


Figure 33: Identification of miR-375 target genes. (A-D) Quantification of the number of occurrences of miR-375 motifs in the 3'UTRs of both up-regulated and down-regulated transcripts in 375KO tissues. Black dots indicate the number of occurrences of miR-375 motifs in the 5% most up- and down-regulated genes (right and left plots, respectively). The distribution of the number of occurrences of motifs complementary to "random miRNAs" in these transcripts is represented as a box plot: blue boxes show the interquartile range, the black whiskers indicate the range of 1.5 times the interquartile range, and the red dots represent the motifs whose number of occurrences falls outside of this range. The "random" miRNAs are selected to have approximately the same number of complementary motifs as miR-375 in the entire set of 3'UTRs. (E) Density plot showing that transcripts with a miR-375 motif (dashed red line) or an evolutionarily-conserved miR-375 motif (solid blue line) are up-regulated in the 375KO relative to transcripts that do not contain this motif.

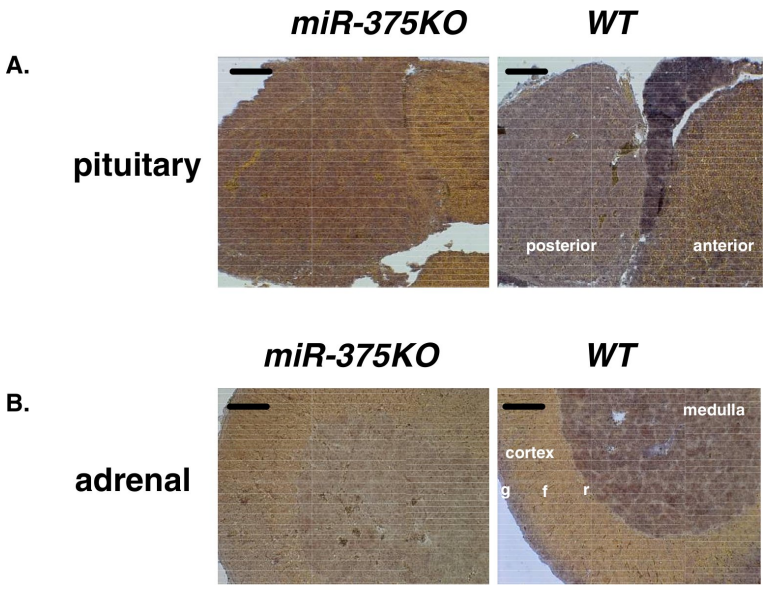


Figure 34: (A) Detection of miR-375 expression by in situ hybridization in pituitary anterior and posterior regions and (B) in adrenal sections from wild-type and 375KO mice using a sequence specific probe for miR-375. miR-375 is detected in the adrenal medulla and the zona glomerulosa (g) of the cortex and not the fasciculata (f) or reticularis (r). (Bar, 50 μm.)

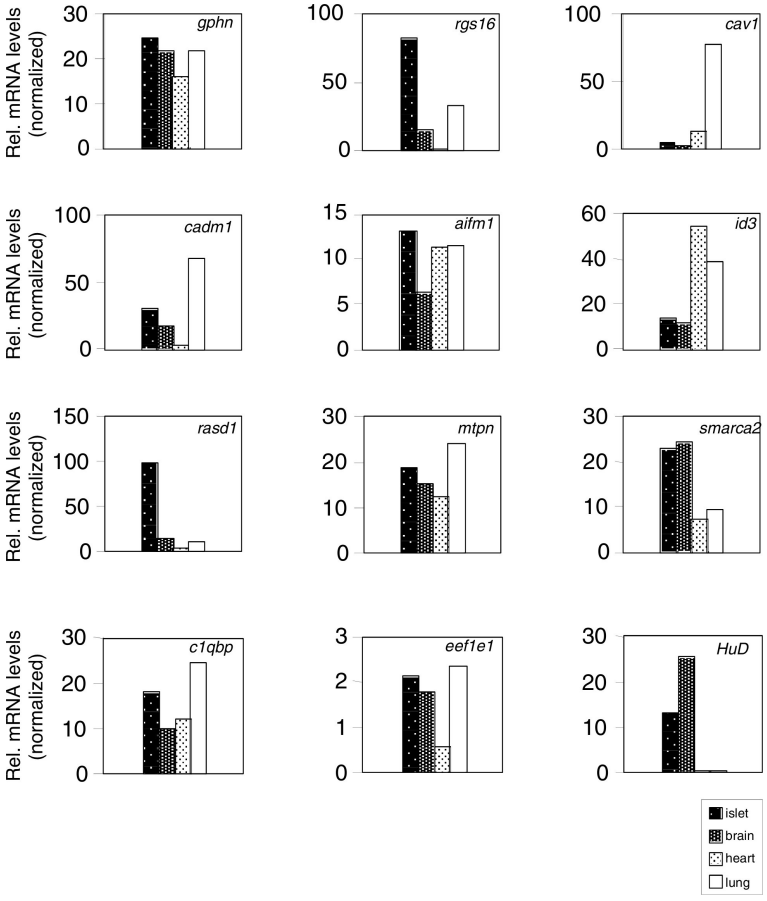


Figure 35: Real-time PCR analysis of miR-375 targets in islets, brain, heart, and lung. Gene expression in the indicated tissues was normalized to U6 levels.

B

SUPPLEMENTARY MATERIAL TO CHAPTER 3

B.1 SUPPLEMENTARY METHODS

B.1.1 *Plasmids and cell culture*

FlPin293 cells stably expressing FLAG/HA EIF2C2 were described in Landthaler et al. [130].

The DNA oligonucleotides used for the amplification of 3′ UTRs were the following (restriction sites are underlined):

CHMP4A-1: 5′-CCGCTCGAGTAAATCTGGGCTTGTCTTCCTAATGCTACC,
CHMP4A-2: 5′-GAATGCGGCCGCGGGAACAAGGGCATTATAACTGCTATCAAAG;
LOC134145-1: 5′-CCGCTCGAGACTTGACTGGGAGTGTTTTTCTGAAATATTGTAG,
LOC134145-2: 5′-GAATGCGGCCGCAAGTTTAGTTAAAGATGTGACCATCTTACTTCATTAC;
AXIN1-1: 5′-CCGCTCGAGCAAAGTGGAGAAGGTGGACTGATAG,
AXIN1-2: 5′-GAATGCGGCCGCTCATTATTATCCAAGTACCTTTGAAAAGATAATTAATTG;
ANXA5-1: 5′-CCGCTCGAGTGTCACGGGGAAGAGCTCCCTG,
ANXA5-2: 5′-GAATGCGGCCGCTCATTAATCTTTTGAATACAATCATCATAATTTTACAGG;
KANK1-1: 5′-CCGCTCGAGTATGCAAATAGCCCTTTATTTACATGCCAC,
KANK1-2: 5′-GAATGCGGCCGCTTTGAAAATATGGCAAGAGTCTAAGGCACTTC;
PGRMC2-1: 5′-CCGCTCGAGACTTTGTAAACAACCAAAGTCAGGGGCCTTC
PGRMC2-2: 5′-GAATGCGGCCGCGTACATGCTTTATTAAAATGGTACTTGTATTTACAG;
RNF128-1: 5′-CCGCTCGAGTCTGTGTAAATAGAAAACTTGAACCATTAGTAATAAC,
RNF128-2: 5′-GAATGCGGCCGCACATTTTATATTTAAAGAGAATCAATACAAATTGGGAC.

B.1.2 *Extraction of positives and negatives from replicated transfection experiments*

For the set of "positives" we wanted to select transcripts that, with high probability, are affected in expression across all experiments in which the expression of a miRNA was perturbed. We therefore developed a probabilistic model that, for each transcript containing one or more miRNA seed matches, uses the expression data from over-expression or knock-down experiments of the corresponding miRNA, to calculate the probabilities that the transcript's expression is affected by the miRNA in each of these experiments.

For the purpose of this model, we define a miRNA seed match as a 7mer or 8mer perfect match to the miRNA seed. We assume that our data consists of K pairs of expression measurements, each corresponding to either a miRNA over-expression or miRNA knock-down experiment, which we will refer to as "contrasts". We will let f_t^k denote the $\log_2$ fold-change of expression of transcript t in contrast k.

DISTRIBUTION OF FOLD-CHANGES FOR NON-TARGETS For our model we first need to calculate, for each contrast k, the probability $P_k(f|-)$ that a transcript that is *not* a target, will have a log fold change of f. To estimate the distributions $P_k(f|-)$ we assumed that they are Gaussian with means μ_k and standard deviation σ_k to be estimated from the data for each contrast k. We in addition assumed that transcripts that do not carry at least a heptameric seed-complementary site

are unlikely to be real targets, and thus estimated μ_k and σ_k from the observed expression changes of transcripts without seed matches.

DISTRIBUTION OF FOLD-CHANGES FOR TARGETS We similarly need to calculate, for each contrast k, a distribution $P_k(f|+)$ that a transcript which is a true target of the miRNA, will have a fold-change f. As little is currently known of the distribution of the severity of the effect that miRNAs have on the expression of their targets we will assume as little as possible about the distribution $P_k(f|+)$. The only thing that we will assume is that a true target must change expression in the right direction, i.e. $f < 0$ for a miRNA over-expression experiment, and $f > 0$ for a miRNA knock-down experiment, and that expression changes are limited to a finite range. That is, we will assign a *uniform* distribution. For example, in the case of a contrast related to a miRNA over-expression:

$$P_k(f|+) = \begin{cases} \frac{1}{|F_k|} & \text{if } F_k \leqslant f < 0 \\ 0 & \text{otherwise} \end{cases}$$

where $F_k = \min_t(f_t^k)$ is the largest negative $\log_2$ fold-change observed in contrast k. The distribution is defined in a similar fashion for contrasts related to a miRNA knock-down, except it is uniform over *positive* instead of negative values.

COMPUTING THE PROBABILITY OF A FUNCTIONALITY PATTERN GIVEN THE DATA The simplest assumption that one could make is that each transcript t is either a true target in each contrast or not a target in any of the contrasts. However, inspection of the data strongly suggested that a transcript t can show a strong response in some experiments and no responses in others. Therefore we developed a more general model in which a transcript can be a "functional target" in some experiments and a non-target in other experiments. We define a functionality pattern α as $\alpha \in S := \{+,-\}^K$. For instance, $\alpha = (\alpha_1, \alpha_2) = (-,+)$ means that the transcript is not a functional target in the first contrast but it is a functional target of the miRNA in the second contrast.

Let D be the whole set of microarray data $D := \{f_t^k\}_{t=\{1,\ldots,T\},k=\{1,\ldots,K\}}$, with T being the number of transcripts and K the number of contrasts. Let further D_t be the microarray data we have about transcript t, $D_t := \{f_t^k\}_{k=\{1,\ldots,K\}}$.

Consider the case where we have $K = 2$ contrasts. What would like to compute ultimately is the posterior probability that a transcript t, which is harboring a seed match (transcripts without seed matches are assumed non-targets per definition), is a functional target of the miRNA whose expression we perturbed in the 2 experiments given the observed $\log_2$ expression fold changes f_t^1, f_t^2. Using Bayes' theorem, we have

$$P(+,+|f_t^1, f_t^2) = \frac{P(f_t^1, f_t^2|+,+)\rho_{++}}{\sum_{\alpha\in S} P(f_t^1, f_t^2|\alpha)\rho_\alpha}.$$

Here we have introduced the *prior* probabilities ρ_α which give the probabilities that a randomly chosen transcript with a seed match will have functionality pattern α. For example ρ_{++} is the prior probability that a randomly chosen transcript with seed match is functional

in both contrasts. As shown below, the ρ_α are unknown parameters which we set by maximizing the probability of the data D.

FITTING PRIOR PROBABILITIES Under our model, the probability of the observed fold-changes D_t for a given transcript t is given by

$$P(D_t|\rho) = \sum_\alpha \rho_\alpha \prod_{k=1}^{K} P_k(f_t^k|\alpha_k) = \sum_\alpha \rho_\alpha P(D_t|\alpha), \tag{B.1}$$

where α_k is the k-th component of the functionality pattern α (either $-$ or $+$), and we have defined the probability $P(D_t|\alpha)$ of the data D_t given pattern α in the last equality. The probability of the entire data set is simply the product over all transcripts t:

$$P(D|\rho) = \prod_{t=1}^{T} \left[\sum_\alpha \rho_\alpha P(D_t|\alpha) \right]. \tag{B.2}$$

We now want to maximize $P(D|\rho)$ with respect to the prior probabilities ρ_α while satisfying the constraint $\sum_\alpha \rho_\alpha = 1$. This can be done using the method of Lagrange multipliers. We let $L(\rho)$ denote the log-likelihood of the parameters ρ, i.e.

$$L(\rho) = \log\left[P(D|\rho)\right]. \tag{B.3}$$

The optimal ρ_α then satisfy the following equations

$$\frac{\partial L(\rho)}{\partial \rho_\alpha} = c \;\forall \alpha, \tag{B.4}$$

where c is a constant (the Lagrange multiplier).

We find for the derivative of the log-likelihood

$$\frac{\partial L(\rho)}{\partial \rho_\alpha} = \sum_{t=1}^{T} \frac{P(D_t|\alpha)}{\sum_\beta \rho_\beta P(D_t|\beta)}. \tag{B.5}$$

From the above equation it is easy to see that

$$\sum_\alpha \rho_\alpha \frac{\partial L(\rho)}{\partial \rho_\alpha} = T. \tag{B.6}$$

Combining this with equation (B.4) we find that the Lagrange multiplier is given by

$$c = T \tag{B.7}$$

and from this it follows that, at the optimum, the ρ_α satisfy:

$$\rho_\alpha = \frac{1}{T} \sum_{t=1}^{T} \frac{P(D_t|\alpha)\rho_\alpha}{\sum_\beta \rho_\beta P(D_t|\beta)}. \tag{B.8}$$

We can solve these equations using an Expectation-Maximization (EM) procedure. We start with a random distribution ρ and use the above equation as an update equation, i.e. at each iteration with replace ρ with $\tilde{\rho}$ according to the equation

$$\tilde{\rho}_\alpha = \frac{1}{T} \sum_{t=1}^{T} \frac{P(D_t|\alpha)\rho_\alpha}{\sum_\beta \rho_\beta P(D_t|\beta)}, \tag{B.9}$$

until the distribution no longer changes. It is easy to show that the second derivatives of the log-likelihood are all negative, i.e.

$$\frac{\partial^2 L(\rho)}{\partial \rho_\alpha \partial \rho_\beta} \leqslant 0 \; \forall \alpha, \beta. \tag{B.10}$$

Therefore, the log-likelihood $L(\rho)$ is a convex function and the EM procedure will lead to the unique global optimum which we will denote by ρ^*.

POSTERIOR PROBABILITIES OF FUNCTIONALITY Using the fitted priors ρ^*_α we can now calculate, for each transcript t, the posterior probabilities $P(\alpha|D_t)$ that it has functionality pattern α. Using Bayes' theorem we have

$$P(\alpha|D_t) = \frac{P(D_t|\alpha)\rho^*_\alpha}{\sum_\beta P(D_t|\beta)\rho^*_\beta}. \tag{B.11}$$

In particular, for the cases where there are two contrasts (like in our data) we can calculate the posterior probabilities $P(++|D_t)$ that the transcript is functional in both contrasts (see Supplementary Figure 50). We sorted all transcripts by this probability $P(++|D_t)$ and selected the positives as the top n transcripts in this list.

NEGATIVES On the other hand, as negatives we wanted to select transcripts that behave consistently, i.e. not responding, in replicated experiments. We therefore computed the sum of squared log2 fold changes of each transcript in the two experiments and we chose a number of transcripts matching the number of positives starting from the lowest to the highest value.

B.2 SUPPLEMENTARY FIGURES

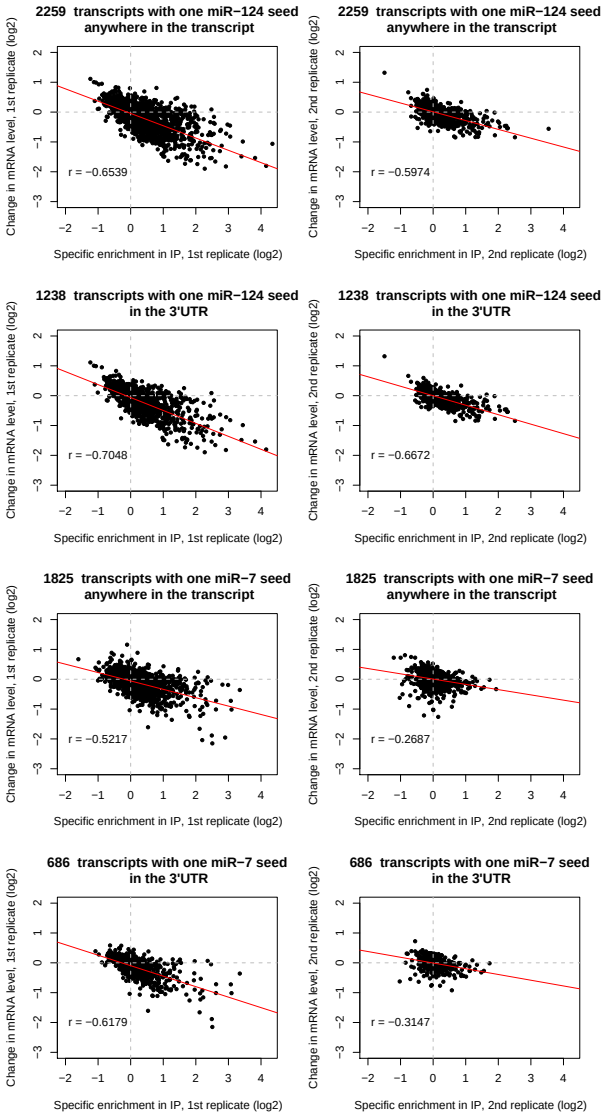


Figure 36: Correlation between the degree of EIF2C2 binding and the extent of mRNA degradation in transcripts in which the single miRNA seed-complementary site is located in the 3'UTR or anywhere in the transcript for the miR-124 and miR-7 EIF2C2-IP. Each row shows a given comparison for the replicate experiments: miR-124 seed match anywhere in the transcript, miR-124 seed match in 3' UTR, miR-7 seed match anywhere in the transcript, miR-7 seed match in 3' UTR. The values of the Pearson correlation coefficients are indicated on the plots and the number of transcripts used for each plot is indicated in the corresponding title.

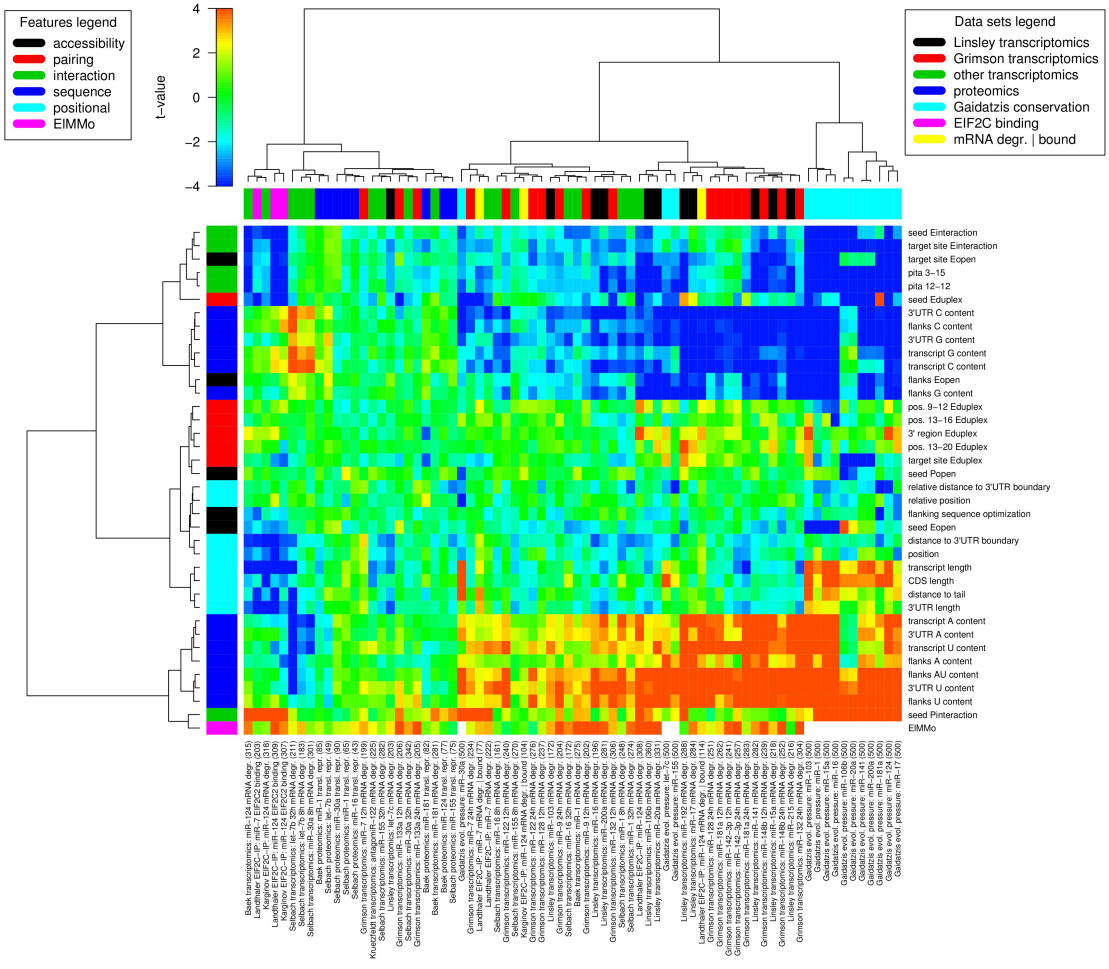


Figure 37: Predictive power of different features of putative miRNA target sites (rows) in predicting functional sites across the 74 data sets (columns). The data sets covered transcriptomics and proteomics measurements after miRNA transfection, transcriptomics measurements after miRNA knock-down, profiling of mRNAs bound to EIF2C/miRNA complexes, and target prediction based on comparative genomics. The column label indicates the source of the data set, the miRNA that was perturbed and the type of measurement that was performed, and the total number of sites involved in the analysis (positives + negatives). The heat-map shows the t-values comparing the distributions of feature values in functional vs non-functional miRNA target sites. The red color indicates positive predictors of miRNA functionality, while the blue color negative predictors of miRNA functionality. The dendrograms of features and data sets were produced through hierarchical clustering using Ward linkage on the euclidean space of t-values. Supplementary Figure 52 in which the same data sets are sorted by the GC-content of the miRNA shows that the target site properties are not miRNA-specific.

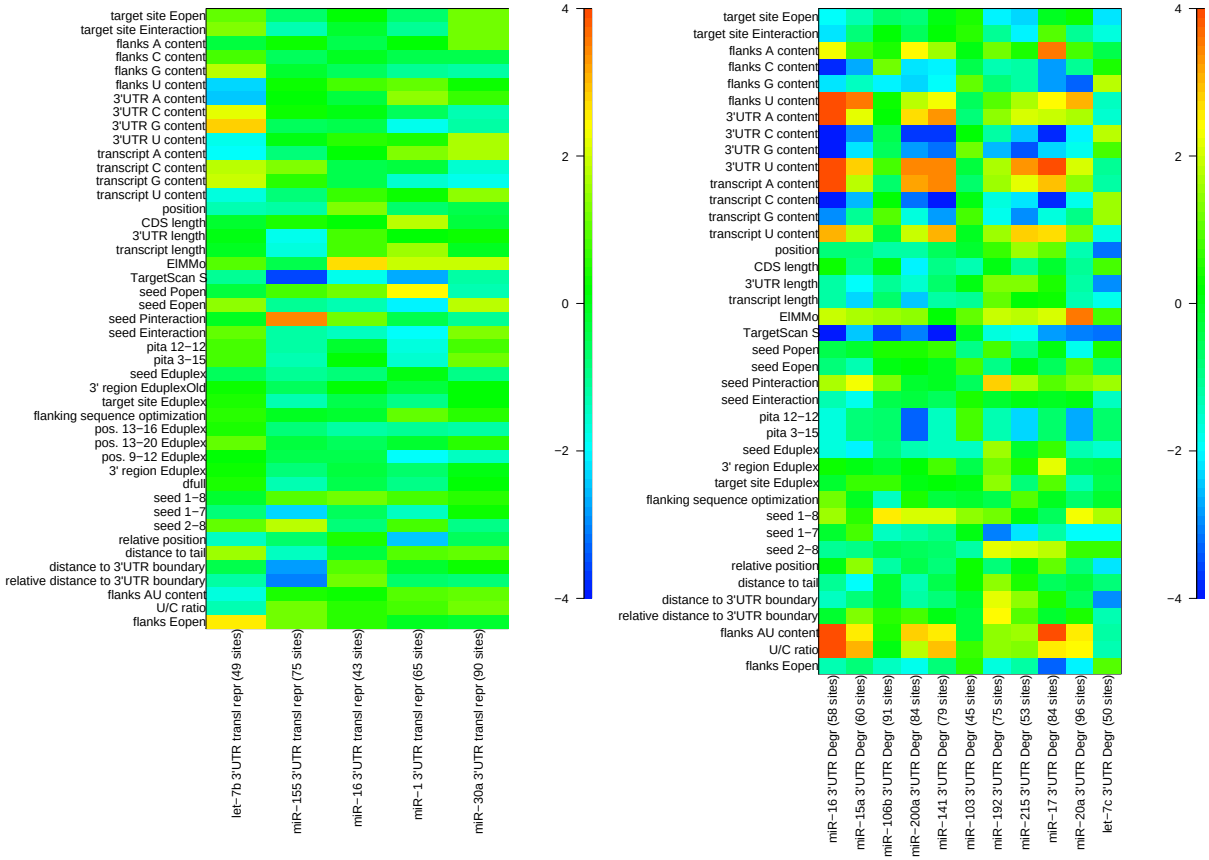


Figure 38: The smaller sample size in the proteomics miRNA transfection experiments cannot, on its own, explain the lack of predictive power that the features that we considered have for the proteomics data. Left panel: detail of the Supplementary Figure 37, showing the predictive power of different features on the proteomics experiments of Selbach et al. [191]. The shot-gun proteomics approach used by the authors (as well as by Baek et al. [8]) makes it possible to quantify the change in concentration of 2000–3000 proteins following the transfection of a miRNA. Except for a few exceptions, most of the features we examined are not predictive of the functionality of miRNA binding sites in this series of experiments. Right panel: we replicated the feature analysis shown on Supplementary Figure 37 using only 2000 randomly selected genes from the miRNA transfection experiments analyzed with microarrays by Linsley et al. [140]. We then determined the predictive power of different features of the miRNA binding sites using the same selection criteria as for the proteomics datasets, i.e. comparing the 75 most down-regulated mRNAs to the 75 least regulated mRNAs. Despite the reduction of the sample size by a factor 3 – 7.5, the predictive power of most sequence features as well as of some structure features is still detectable in most experiments. Therefore, the sample size cannot explain on its own why none of features we study appear to be predictive of the miRNA binding sites identified by the proteomics experiments.

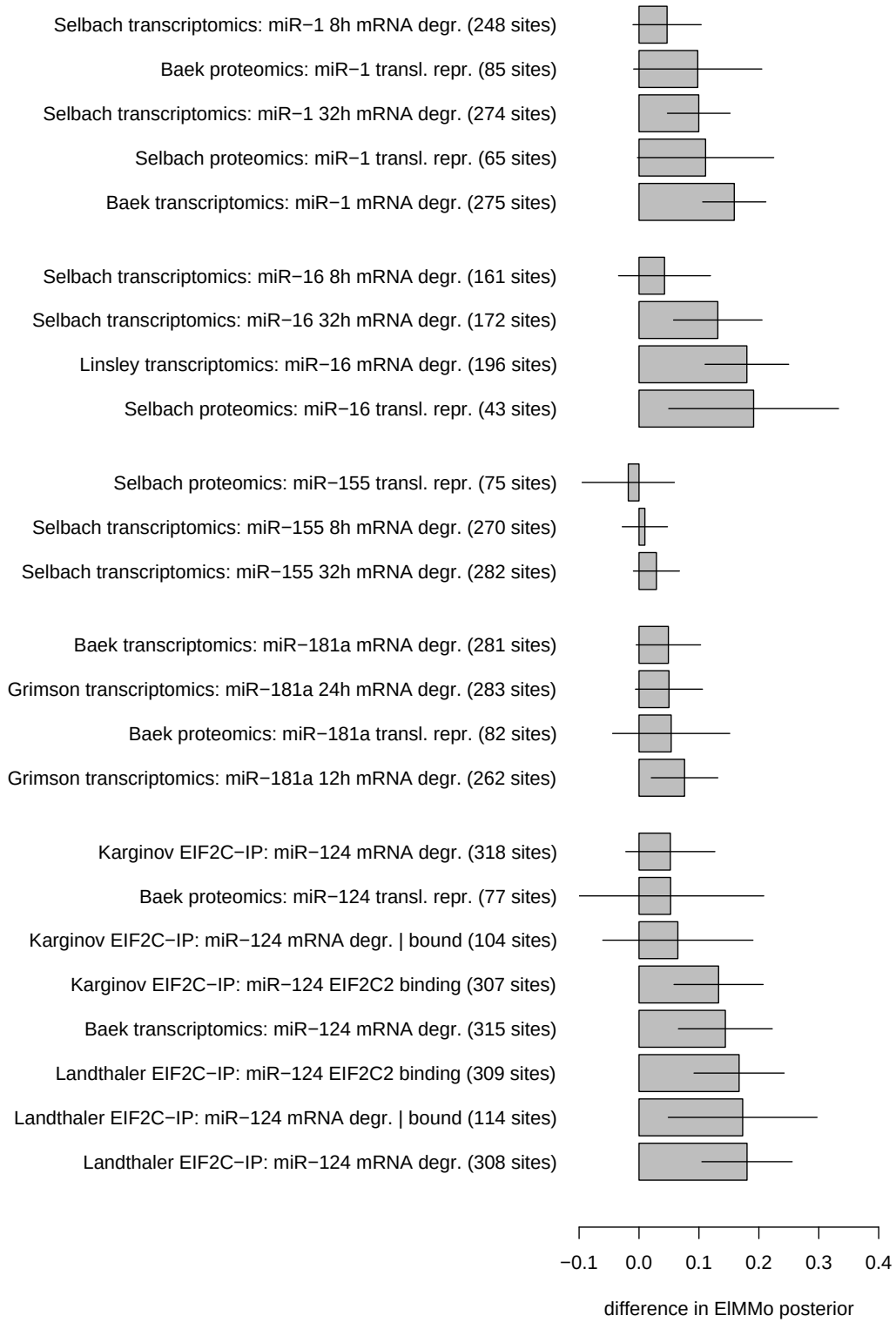


Figure 39: Difference between the average ElMMo posterior of functional vs non-functional miRNA target sites in different experiments.

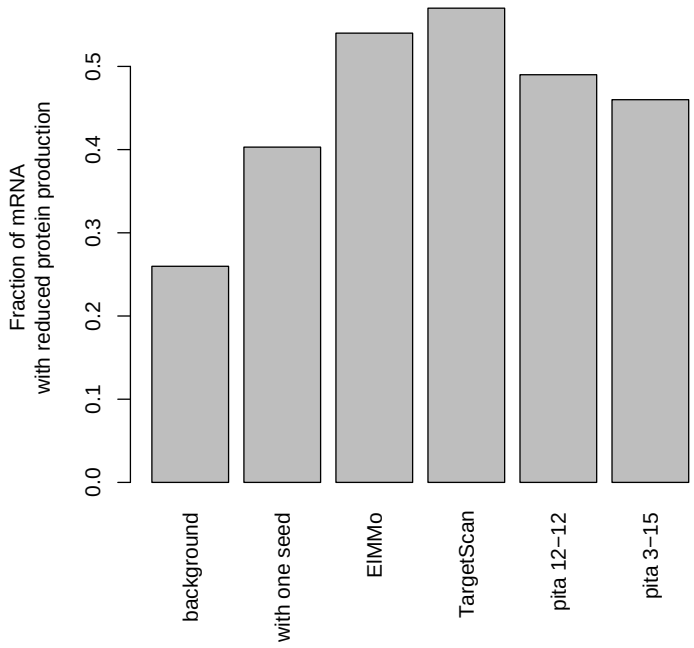


Figure 40: Fraction of the mRNAs obtained by applying a given "prediction" method that have reduced protein production according to the pSILAC experiments of Selbach et al. [191].

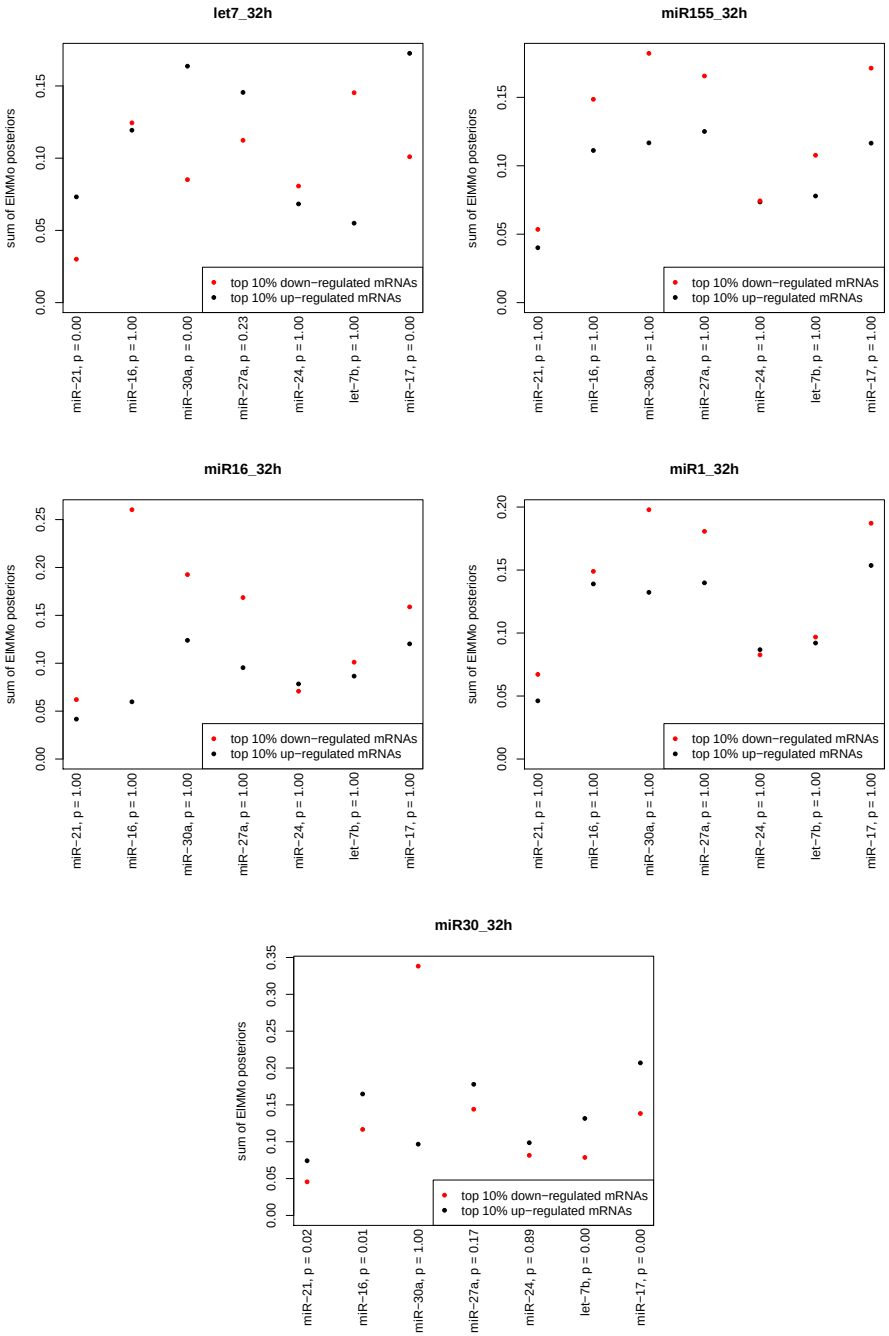


Figure 41: Expected number of evolutionarily selected binding sites for the 7 most abundant miRNAs in HeLa cells [129] in the 10% most up-regulated and down-regulated transcripts in individual transfection experiments of Selbach et al. [191]. The expected number of sites were computed by summing the ElMMo posteriors over all putative binding sites for a miRNA within every 3'UTR. Each panel represents one transfection experiment, where the transfected miRNA is indicated in the title. The expected number of binding sites in up- and down-regulated transcripts where compared using Wilcoxon's ranks sum test. The corresponding p-values were computed under the alternative hypothesis that up-regulated transcripts harbor more miRNA binding sites under evolutionary pressure than down-regulated transcripts and were corrected for multiple testing using the Bonferroni method.

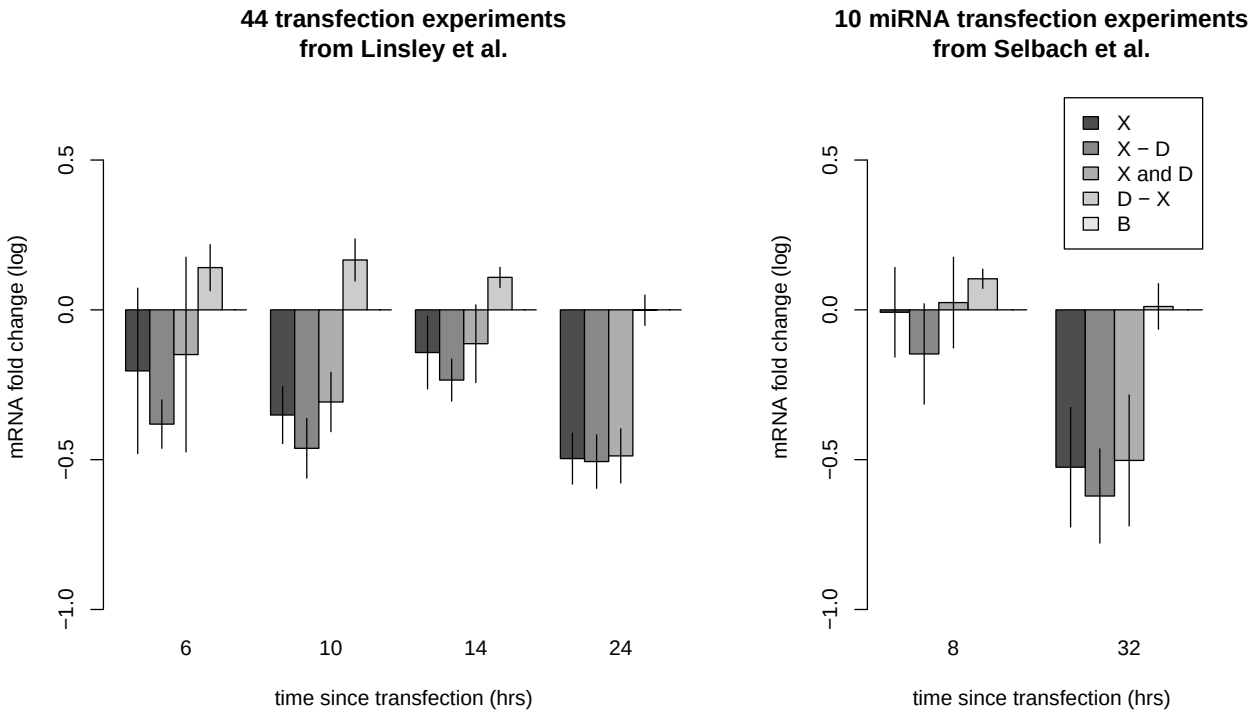


Figure 42: The competition between endogenous miRNAs and the transfected miRNA is transient in time. The y-axis shows average the log fold change of mRNAs carrying seed matches to the transfected miRNA in their 3′ UTR (X), seed matches to the transfected miRNA but not to the most expressed endogenous miRNAs (X - D), seed matches to both the transfected miRNA and the top expressed endogenous miRNAs (X and D), seed matches to the most expressed endogenous miRNAs but not the transfected miRNA (D - X), and seed matches to neither the transfected miRNA nor the endogenous miRNAs (B). The error bars show the 95% confidence interval on the mean after averaging over all miRNA transfection experiments performed at the same time point. Left: re-analysis of 44 microarray experiments performed 6, 10, 14 and 24 hours after miRNA / siRNA transfection in HCT116 Dicer -/- cells [140]. Right: re-analysis of 10 microarray experiments performed 8 and 24h after miRNA transfection in HeLa cells [191].

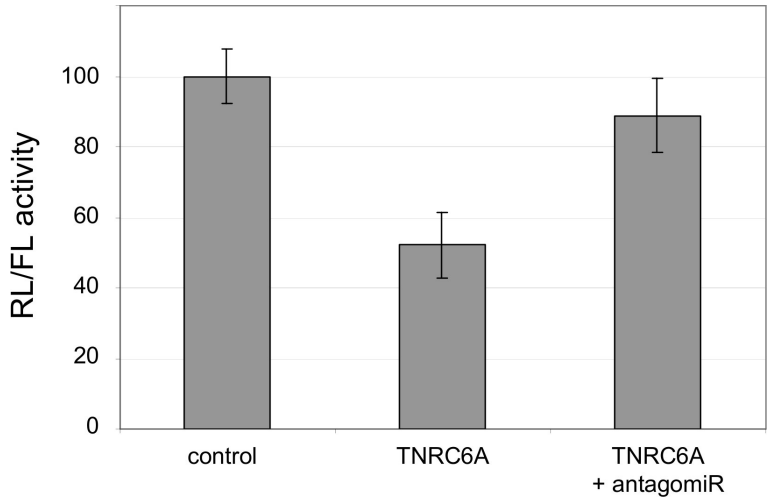


Figure 43: Luciferase reporter assay confirming that *TNRC6A* (also known as *GW182*) is a direct target of the endogenously expressed miR-30a in HeLa cells. The *TNRC6A* 3′UTR was cloned downstream of the coding region of the *Renilla* luciferase (RL) and the vector system subsequently transfected into HeLa cells, either alone (*TNRC6A*) or together with the miR-30a antisense inhibitor (*TNRC6A* + antagomiR). The y-axis shows the change in *Renilla* luciferase activity normalized to the firefly luciferase (FL, control).

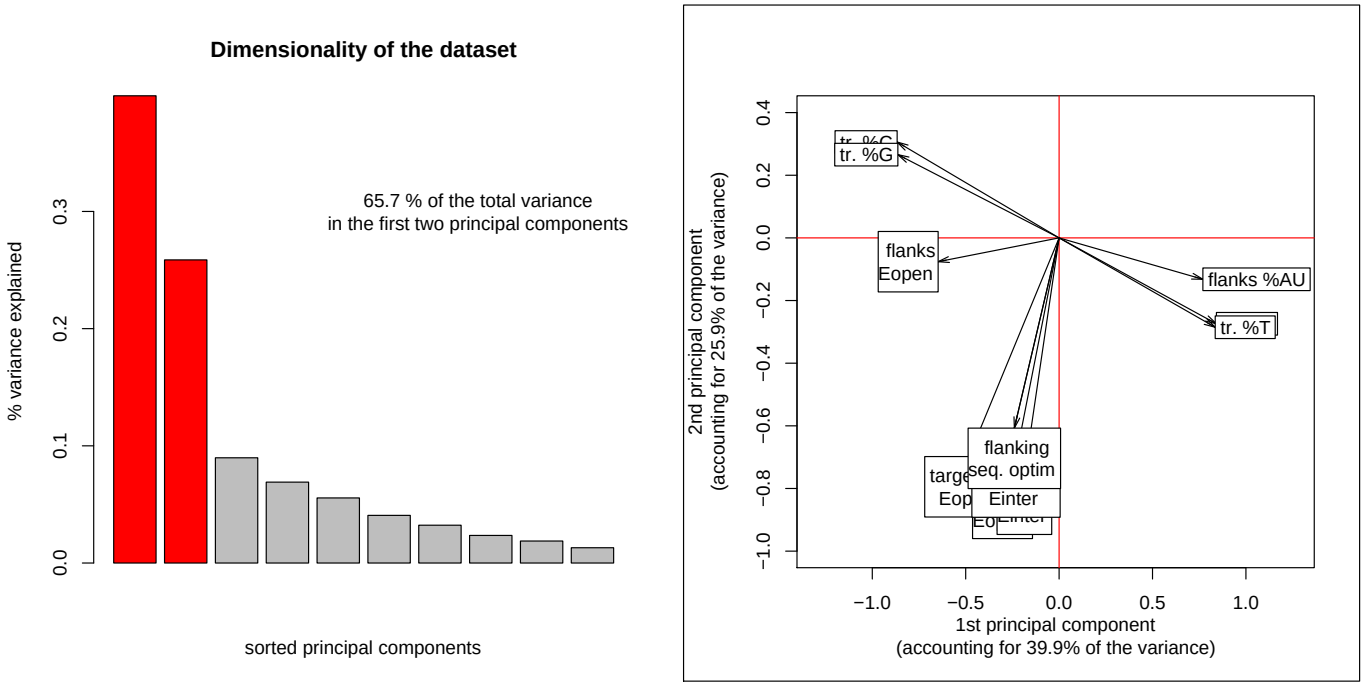


Figure 44: Principal component analysis of a subset of features computed over 5964 miRNA binding sites (positives and negatives) from the comparative genomics data set. Although Figure 6 shows that, when considered independently, both structure as well as sequence features are predictive of miRNA target site functionality, it does not show to what extent these features are redundant. To determine this, we collected all 5964 miRNA binding sites from the comparative genomics dataset (comprising positives as well as negatives) and considered the following set of features: transcript A, G, C and U content, flanks AU content, the seed, target site and flanks Eopen, flanking sequence optimization, seed and target site Einteraction. We then centered and rescaled this subset of features and determined how many principal components are needed to describe this subset of features. The first two principal components accounted for 65.7% of the total variance, with the third component and next components accounting for a substantially smaller amount of the variance compared to the first two principal components (left panel). We then projected the subset of features onto the plane spanned by the first two principal components and determined that sequence features clustered well with the first principal component, while all structure features except "flanks Eopen" clustered together with the second principal component (right panel). This suggests that, except for "flanks Eopen" which correlates with the G and C content, sequence and structure features are not redundant and characterize miRNA binding sites in a complementary way. Performing the same analysis on the smaller transcriptomics and proteomics datasets yielded the same results.

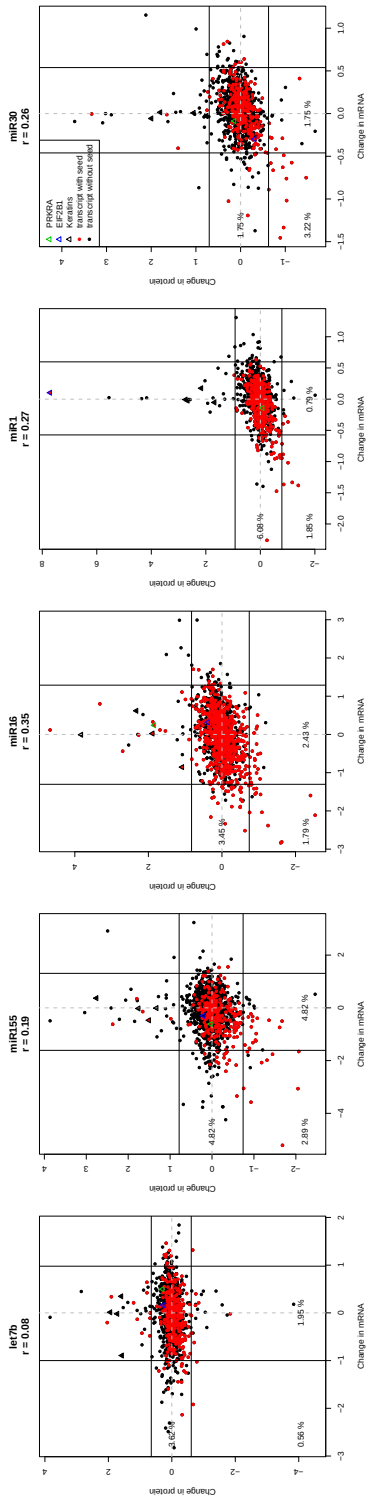


Figure 45: Correlation between change in protein and mRNA levels in the let-7, miR-155, miR-16, miR-1 and miR-30a pSILAC experiments of Selbach et al. [191]. The x-axis shows the log2 fold change in expression between miRNA-transfected to mock-transfected HeLa cells. The black lines indicate the cut-offs in mRNA and protein level fold change beyond which we consider the mRNA or the protein differentially expressed. Red and black dots respectively represent transcripts that carry at least one match or do not carry any match to the seed of the transfected miRNA. The three percentages respectively indicate the proportion of transcripts carrying at least on miRNA seed match that are down-regulated at the protein level only, at the mRNA level only, or both at the levels of the protein and mRNA. r is the Pearson correlation coefficient between the change in protein and mRNA levels.

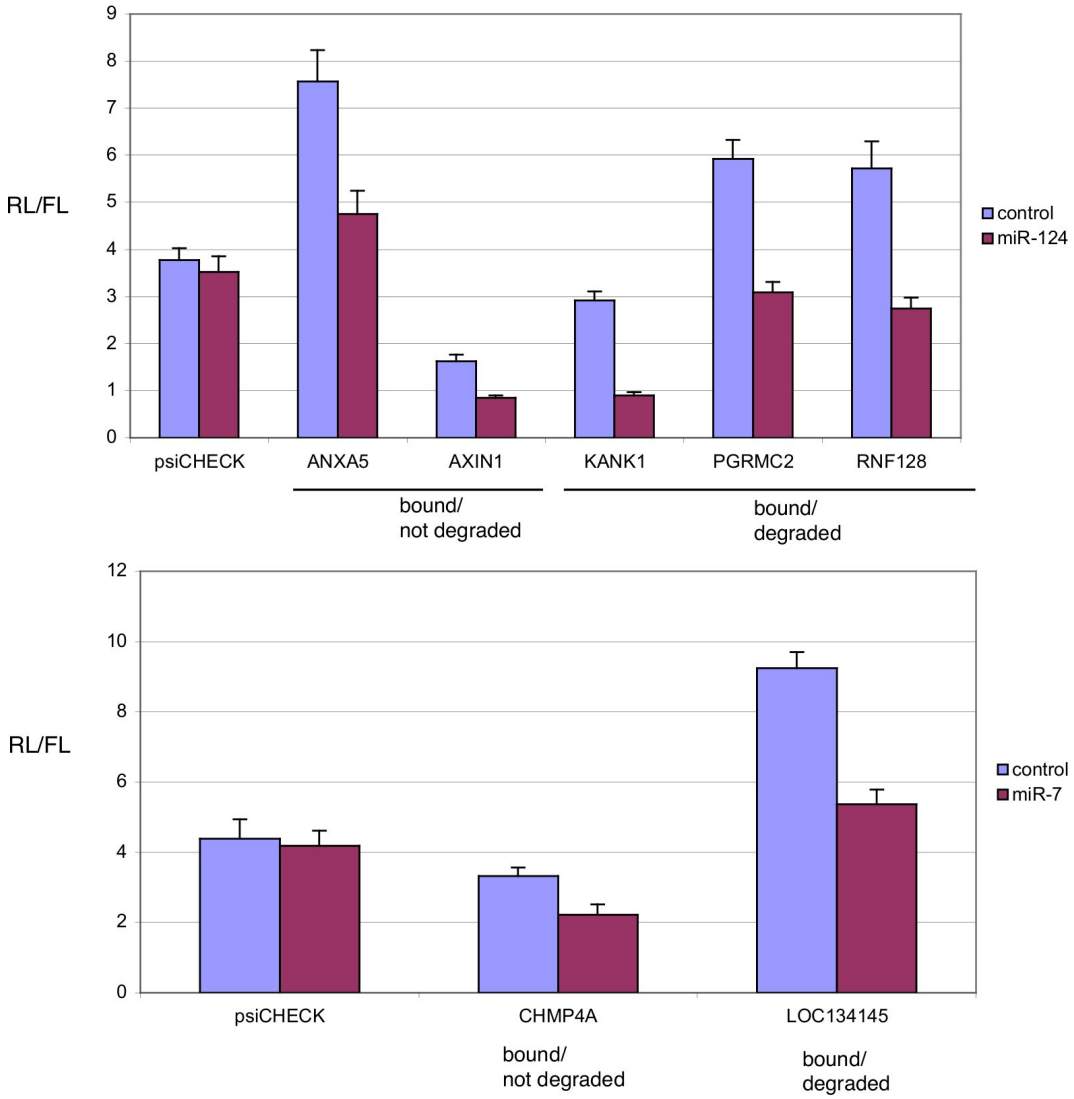


Figure 46: miR-124 and miR-7-mediated repression of 3′UTRs fused to luciferase reporter genes. psiCHECK reporter constructs were generated by fusing the full-length 3′UTRs of the genes indicated to the *Renilla* Luciferase. Dual Luciferase activity from HEK293 cells cotransfected with each reporter psiCHECK construct and miR-124 or miR-7 duplex was compared to cotransfection of each reporter construct with control RNA duplex. Transfections of parental psiCHECK vector without inserted 3′UTR (psiCHECK) is shown. *Renilla* Luciferase versus firefly luciferase activities are indicated. Error bars represent standard deviation computed over 10 replicates.

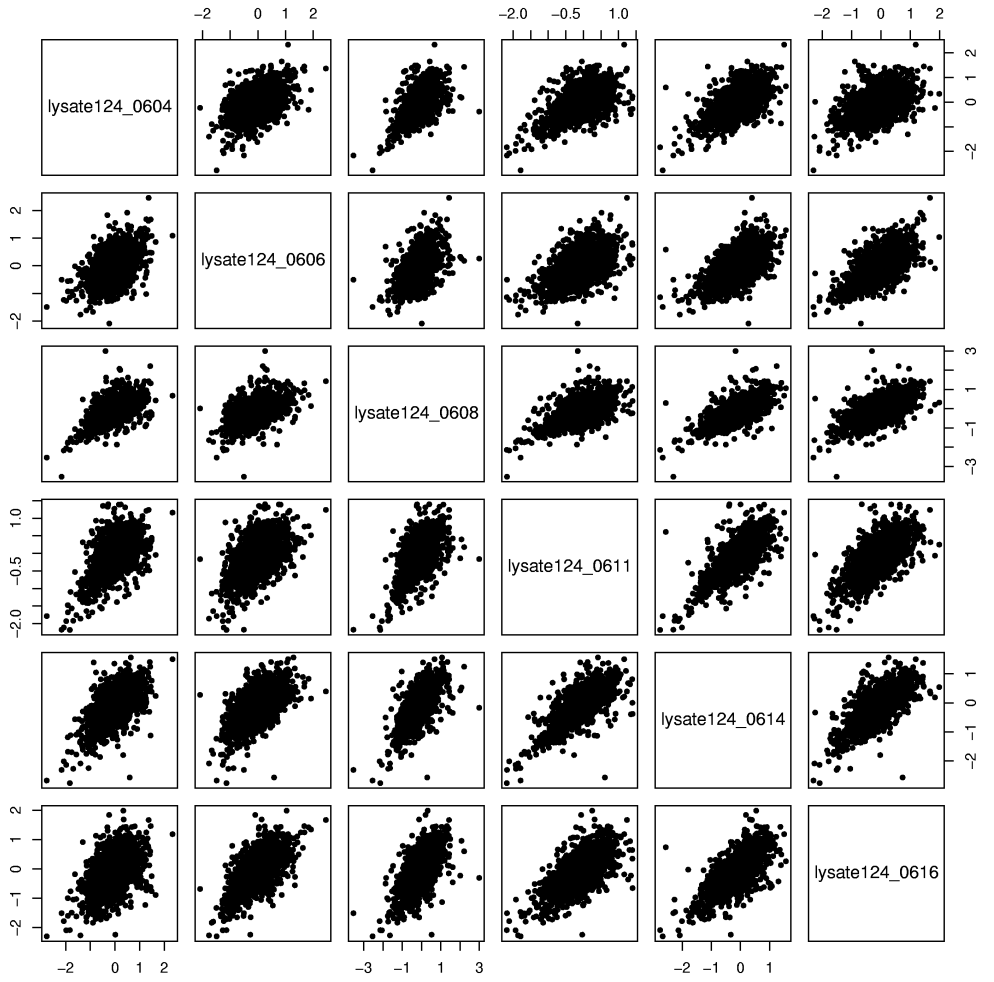


Figure 47: Correlation between the extent of mRNA degradation following miR-124 transfection in the 6 biological replicates of the transcripts of the Karginov et al. EIF2C2-IP dataset. The axes show log10 fold changes in pairs of replicates. The Pearson correlation coefficient between log10 fold changes of replicates ranges from 0.44 to 0.76 depending on the pair of experiments being considered.

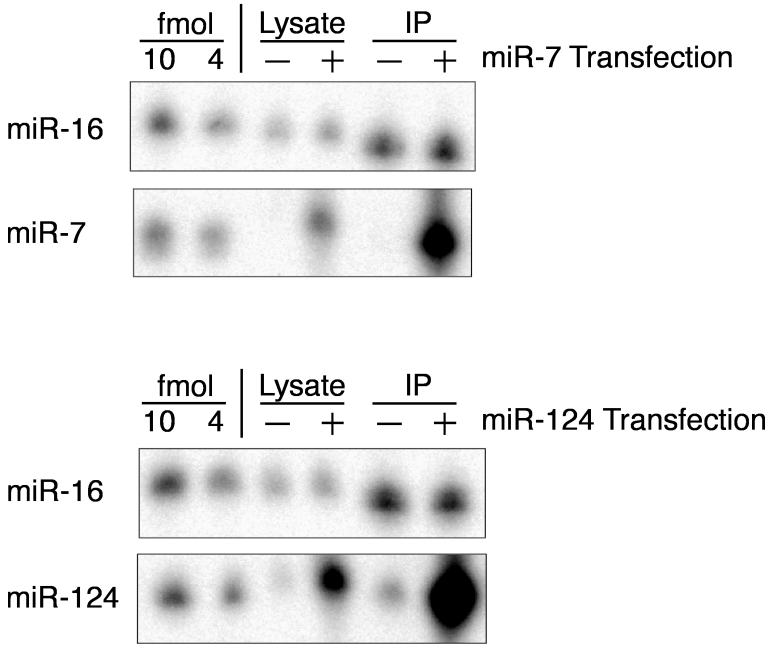
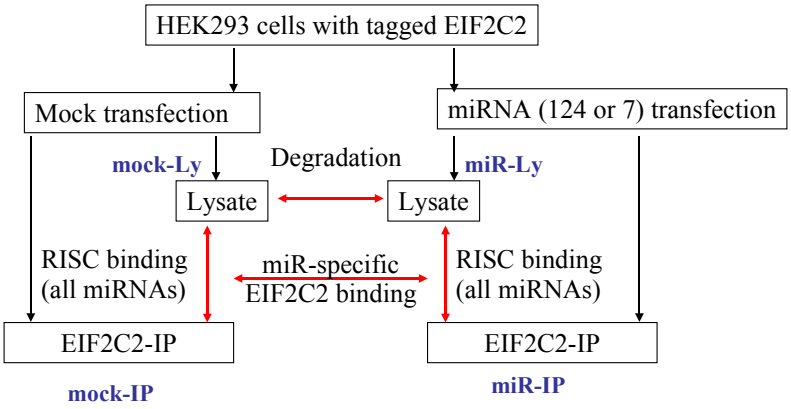


Figure 48: miRNA transfection and immunoprecipitation. Cells stably expressing FLAG/HA-EIF2C2 were mock-transfected (-) and transfected with a miR-7/miR-7* and miR-124/miR-124* duplex (+), respectively. 15 hours after transfection cells were lysed and the epitope-tagged protein was immunoprecipitated from cytoplasmic extracts with FLAG-antibody. RNA was extracted from the cleared cell lysate and the immunoprecipitate (IP). 15 µg total cellular RNA and one fifth of IPed RNA was separated on a 12% polyacrylamide gel, blotted, and probed for miR-16, miR-7, and miR-124, respectively. 10 and 4 femtomole (fmol) of synthetic miR-16, miR-7, and miR-124, were loaded as standards.



Degradation:	miR-Ly / mock-Ly
EIF2C2-binding miRNA:	miR-IP / miR-Ly
EIF2C2-binding mock:	mock-IP / mock-Ly
miR-specific EIF2C2-binding:	(miR-IP / miR-Ly) / (mock-IP / mock-Ly)

Figure 49: Sketch of the computation of the binding and degradation measures: EIF2C2-binding in miRNA transfection is given by the ratio of transcript levels in the immunoprecipitate and in the lysate of miR-transfected cells (miR-IP/miR-Ly); EIF2C2-binding in mock-transfection is given by the ratio of transcript levels in the immunoprecipitate and in the lysate of mock-transfected cells (mock-IP/mock-Ly); miR-specific EIF2C2-binding is given by the ratio of the previous two ratios, (miR-IP/miR-Ly)/(mock-IP/mock-Ly).

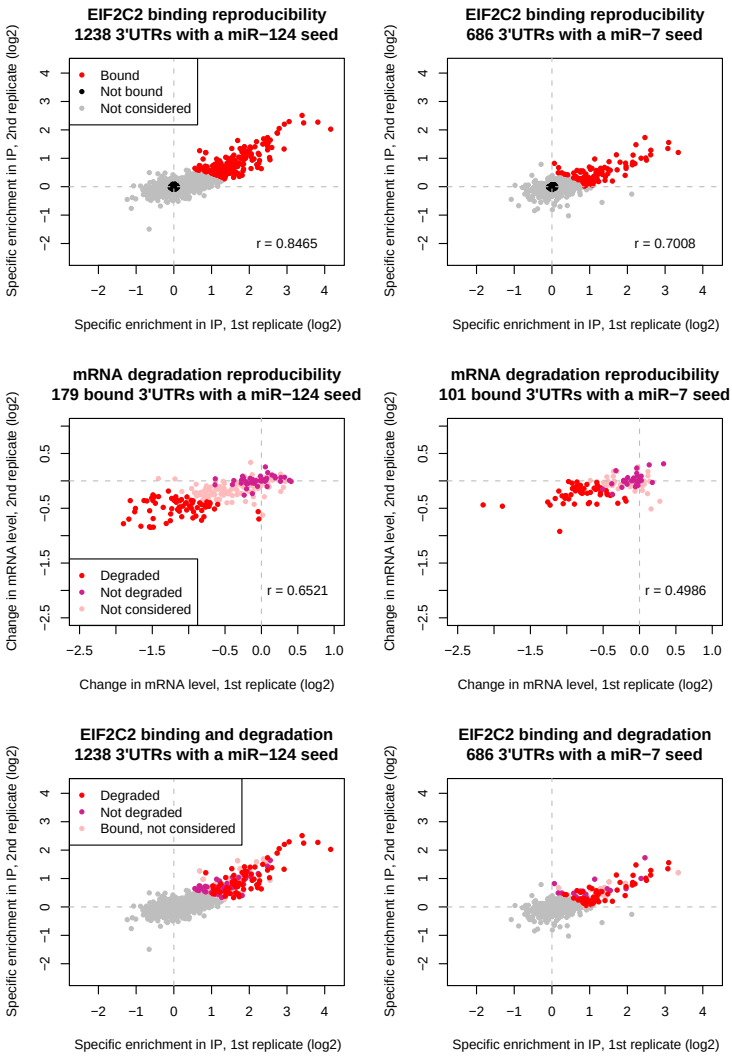


Figure 50: Selection of positive and negative examples for EIF2C2 binding and mRNA degradation upon miR-124 (left) and miR-7 (right) transfection. Upper panels show the correlation between EIF2C2 binding measures in the replicate experiments. The transcripts marked with red were considered "bound" and those marked in black "not bound". The procedure for this selection is described in the supplementary text. Middle panels show the correlation between degradation of bound transcripts (in red in the upper panels) in replicate experiments. Transcripts marked in red were considered "bound and degraded", those in violet "bound but not degraded". Lower panels reproduce the upper panels, except that transcripts that were considered "bound" are further shown in the color that indicates whether they were or not also considered degraded. These figures show that the degree of degradation is not simply proportional to the degree of EIF2C2 binding.

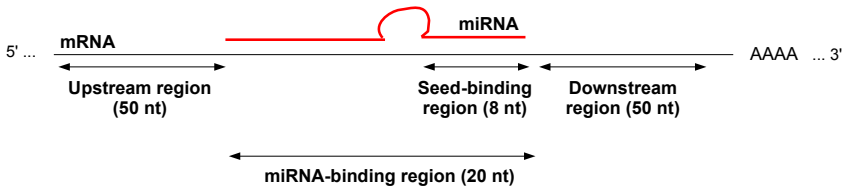


Figure 51: Sketch of the transcript regions used in the computation of structural and sequence features.

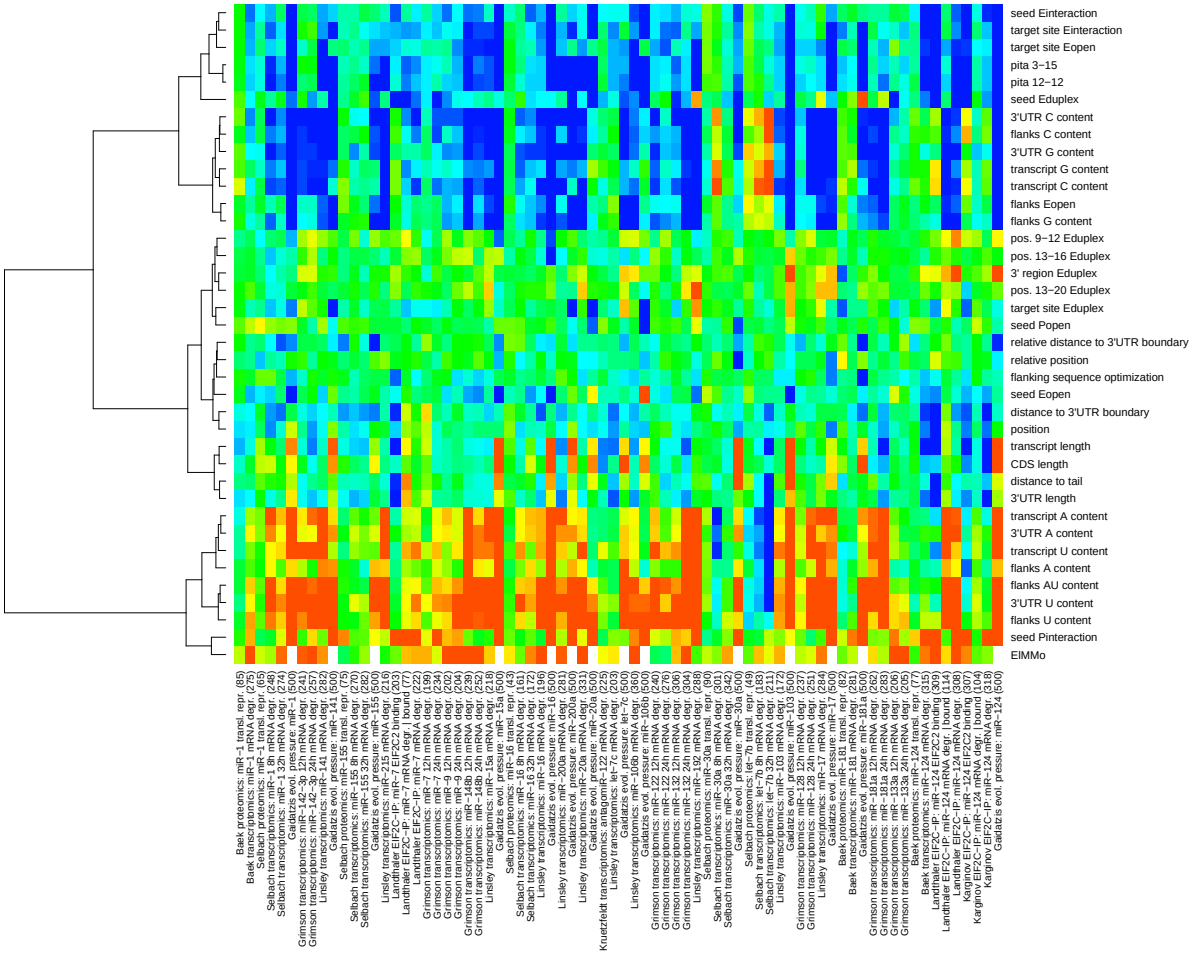


Figure 52: The features predictive of miRNA targeting are not determined by the GC content of the mature miRNA. The present figure shows a heat map similar to ones shown on Figure 5 and Supplementary Figure 37, but in which we reordered the columns (data sets) according to the GC content of the transfected miRNA. The left-most columns correspond to GC-poor miRNAs while the right-most columns feature data sets involving GC-rich miRNAs.

C

SUPPLEMENTARY MATERIAL TO CHAPTER 4

C.1 SUPPLEMENTARY FIGURES

C.2 SUPPLEMENTARY TABLES

Supplementary tables S1 to S7 are available in the online supplementary material of Hafner et al. [90].

C.3 SUPPLEMENTARY EXPERIMENTAL PROCEDURES

Oligonucleotides and siRNA duplexes

The following oligodeoxynucleotides were used for PCR and cDNA cloning into pENTR4 (Invitrogen), restriction sites are underlined:
PUM2, ATGAATCATGATTTTCAAGCTCTTGCATTAG,
ATAAGAATGCGGCCGCTTACAGCATTCCATTTGGTGGTCCTCCAATAG;
QKI, ACGCGTCGACATGGTCGGGGAAATGGAAACG,
ATAAGAATGCGGCCGCTTAGCCTTTCGTTGGGAAAGCC;
IGF2BP1, ACGCGTCGACATGAACAAGCTTTACATCGGCAACCTC,
ATAAGAATGCGGCCGCTCACTTCCTCCGTGCCTGGGCCTG;
IGF2BP2, ACGCGTCGACATGATGAACAAGCTTTACATCGGGAAC,
ATAAGAATGCGGCCGCTCACTTGCTGCGCTGTGAGGCGAC;
IGF2BP3, ACGCGTCGACATGAACAAACTGTATATCGGAAACCTCAG,
ATAAGAATGCGGCCGCTTACTTCCGTCTTGACTGAGGTGGTC;
The following oligoribonucleotides were used for QKI protein in vitro binding and crosslinking studies and were purchased from Dharmacon:

```
GUAUGCCAUUAACAAAUUCAUUAACAA
G(4SU)AUGCCAUUAACAAAUUCAUUAACAA
GUA(4SU)GCCAUUAACAAAUUCAUUAACAA
GUAUGCCA(4SU)AACAAAUUCAUUAACAA
GUAUGCCAU(4SU)AACAAAUUCAUUAACAA
4SU, 4-thiouridine.
```

The following siRNA duplexes (sense/antisense) were used for knockdown experiments and synthesized on a modified ABI 392 RNA/DNA synthesizer using Dharmacon synthesis reagents.
QKI duplex 1, GAAGAGAGCAGUUGAAGAAUU,
UUCUUCAACUGCUCUCUUCUU;
QKI duplex 2, CCAAUUGGGAGCAUCUAAAUdT,
UUUAGAUGCUCCCAAUUGGUdT;
IGF2BP1, GGGAAGAAUCUAUGGCAAAUU,
UUUGCCAUAGAUUCUUCCCUU;
IGF2BP2, GGCAUCAGUUUGAGAACUAUU,
UAGUUCUCAAACUGAUGCCUU;
IGF2BP3, AAAUCGAUGUCCACCGUAAUU,
UUACGGUGGACAUCGAUUUUU.

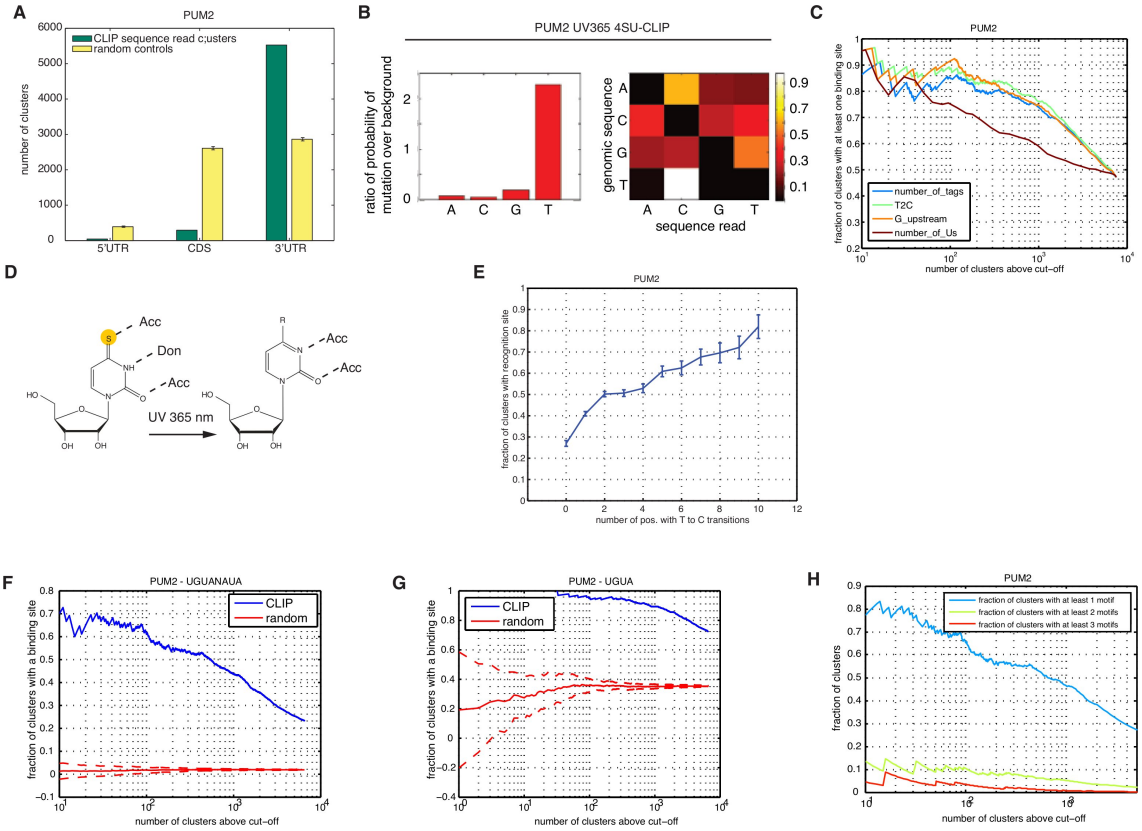


Figure 53: Analysis of PUM2-PAR-CLIP clusters. Related to Figure 12. (A) Analysis of the transcript regional preferences and the mutational pattern of crosslinked sequences of PUM2. The number of exonic sequence read clusters annotated as derived from the 5'UTR, CDS or 3'UTR of a target transcript is shown (green bars). Yellow bars show the expected location distribution of clusters if PUM2 binds without regional preference to the set of target transcripts. (B) Mutational pattern observed with 4SU-PAR-CLIP for PUM2. The left panel indicates the mutation frequency of each of the four nucleotides relative to the frequency of occurrence of these nucleotides in all sequence reads; the right panel shows, for each of the four nucleotides, the frequency of mutation towards each of the three others. In the right panels, white indicates high mutation frequency towards a particular nucleotide. 4SU-PAR-CLIP yields about a 15-fold increased mutation preference for T, nearly always to C. (C) Fraction of clusters containing the PUM2-recognition motif, versus the total number of clusters above a given cut-off on a particular property as indicated in each figure legend (G upstream: number of sequence reads with a G at position -1; T to C: number of sequence reads with a T to C mutation; number of sequences: total number of sequence sequence reads in the cluster, number_of_Us: number of uridines in the sequence read cluster). For each cut-off on a given property, the fraction of clusters with at least one binding site above the given cut-off is shown. Cut-off increases from right to left. The best signal is obtained by sorting according to the frequency of crosslinking events. (D) The increase in T to C transitions after 4SU-protein crosslinking can be rationalized by structural changes in donor/acceptor properties of 4SU after crosslinking to proximal amino acid side chains and subsequent incorporation of dG rather than dA in the reverse transcription; R representing a side chain. (E) Fraction of clusters with the recognition element (as indicated) for PUM2 versus the number of distinct crosslinking sites within a cluster indicated by a T to C change. The fraction of sites containing at least one recognition motif rises with the number of crosslinking sites. (F-H) Enrichment of binding motifs for PUM2 for the consensus motif UGUANAUA (F) as well as the short variant UGUA (G) compared to CCRs with randomized sequences. Panel (H) shows the fraction of clusters with at least one, two or three UGUANAUA motifs. Most clusters contain only one binding site.

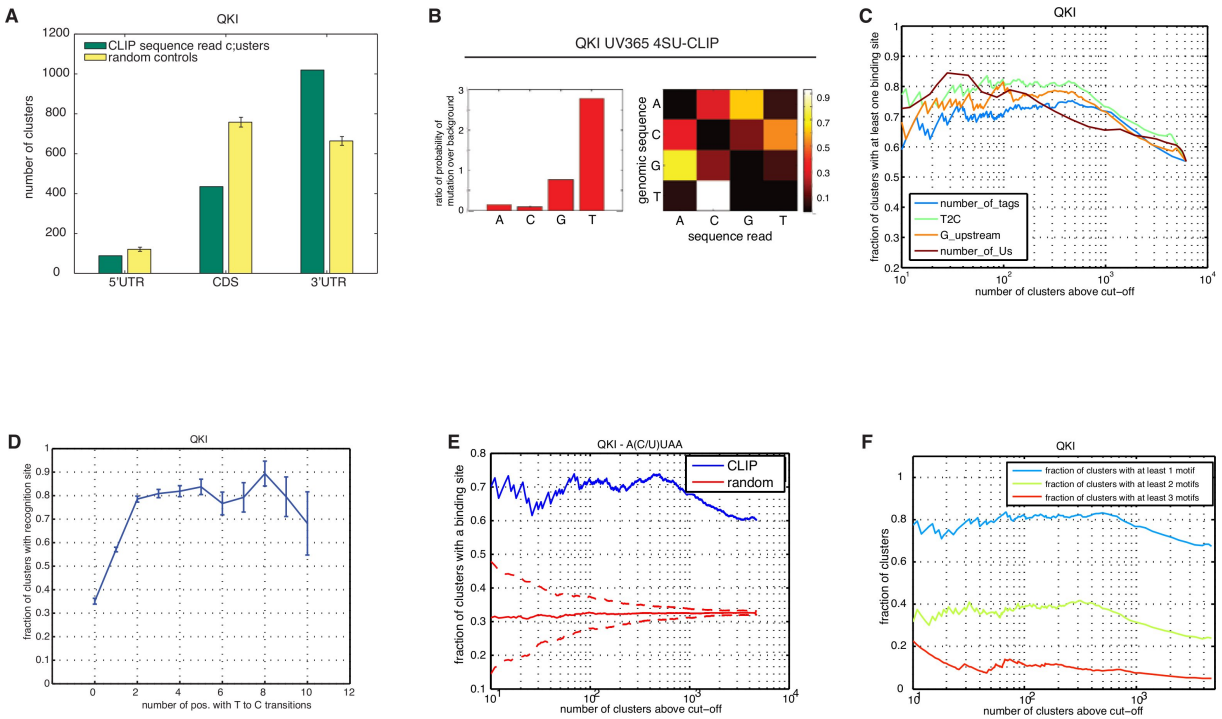


Figure 54: Analysis of QKI-PAR-CLIP clusters. Related to Figure 55. (A) Analysis of the transcript regional preferences and the mutational pattern of crosslinked sequences of QKI. The number of exonic sequence read clusters annotated as derived from the 5'UTR, CDS or 3'UTR of a target transcript is shown (green bars). Yellow bars show the expected location distribution of clusters if QKI binds without regional preference to the set of target transcripts. (B) Mutational pattern observed with 4SU-PAR-CLIP for QKI. The left panel indicates the mutation frequency of each of the four nucleotides relative to the frequency of occurrence of these nucleotides in all sequence reads; the right panel shows, for each of the four nucleotides, the frequency of mutation towards each of the three others. In the right panels, white indicates high mutation frequency towards a particular nucleotide. 4SU-PAR-CLIP yields about a 6-fold increased mutation preference for T, nearly always to C. (C) Fraction of clusters containing the PUM2-recognition motif, versus the total number of clusters above a given cut-off on a particular property as indicated in each figure legend (G upstream: number of sequence reads with a G at position -1; T to C: number of sequence reads with a T to C mutation; number of sequences: total number of sequence sequence reads in the cluster, number_of_Us: number of uridines in the sequence read cluster). For each cut-off on a given property, the fraction of clusters with at least one binding site above the given cut-off is shown. Cut-off increases from right to left. The best signal is obtained by sorting according to the frequency of crosslinking events. (D) Fraction of clusters with the recognition element (as indicated) for QKI versus the number of distinct crosslinking sites within a cluster indicated by a T to C change. The fraction of sites containing at least one recognition motif rises with the number of crosslinking sites. (E) Enrichment of the A(C/U)UAA binding motif in CCRs of QKI. Panel (F) shows the fraction of clusters with at least one, two or three motifs. A significant fraction of clusters contains two or more binding sites.

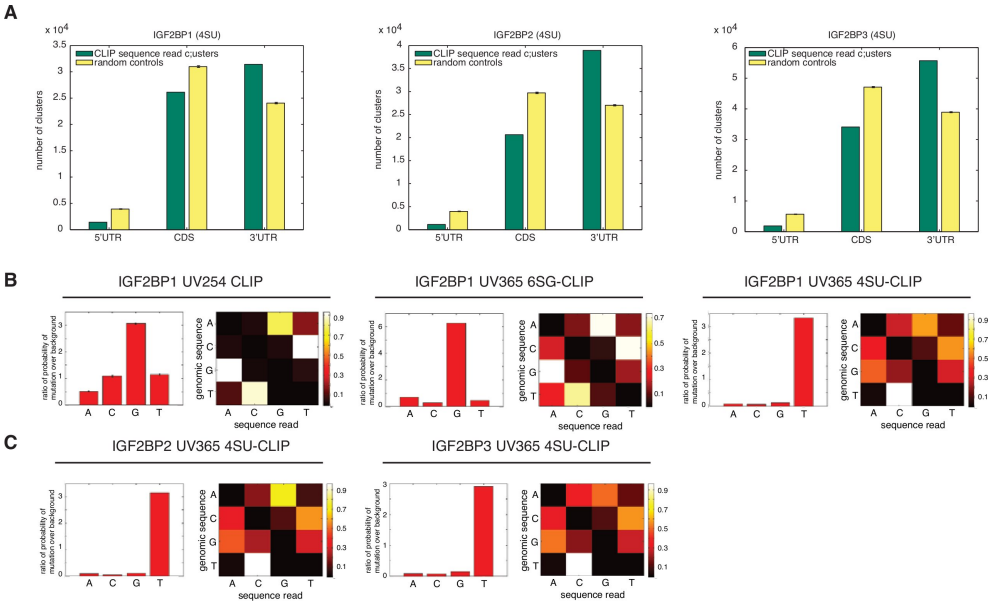


Figure 55: Analysis of IGF2BP1-3-PAR-CLIP clusters. Related to Figure 14. (A) Analysis of the transcript regional preferences and the mutational pattern of crosslinked sequences of IGF2BP1-3. The number of exonic sequence read clusters annotated as derived from the 5'UTR, CDS or 3'UTR of a target transcript is shown (green bars). Yellow bars show the expected location distribution of clusters if IGF2BP1-3 bind without regional preference to the set of target transcripts. (B) Comparison of the mutational patterns observed with traditional UV 254 nm CLIP of HEK293 cells stably expressing FLAG/HA-tagged IGF2BP1 and that observed with UV 365 nm CLIP of cells grown in 6SG or 4SU containing medium. For each experimental condition two panels are shown: the left one indicates the mutation frequency of each of the four nucleotides relative to the frequency of occurrence of these nucleotides in all sequence reads; the right one shows, for each of the four nucleotides, the frequency of mutation towards each of the three others. In the right panels, white indicates high mutation frequency towards a particular nucleotide. In general, transitions are more frequent than other mutation types. Traditional 254 nm CLIP generates mutations preferably on Gs (left panel). Mutations after UV254 CLIP were twice as frequent at G compared to any other position (left panel) and predominantly identified as G to A transition (shown by the matrix in the right panel). Treatment of cells with 6SG (middle two panels, top row) resulted in a marked preference for mutations at G, about one order of magnitude compared to the other nucleotides with a preferred substitution of the G with an A. The preference for mutations at G is much more pronounced relative to that observed in the 254 nm crosslinked cells. 4SU-CLIP yields about a 30-fold increased mutation preference for T, nearly always to C. (C) Same analysis as in (B) for IGF2BP2 and 3. The mutational biases for these proteins are comparable. T is almost exclusively targeted for mutation, and is preferentially sequenced as C.

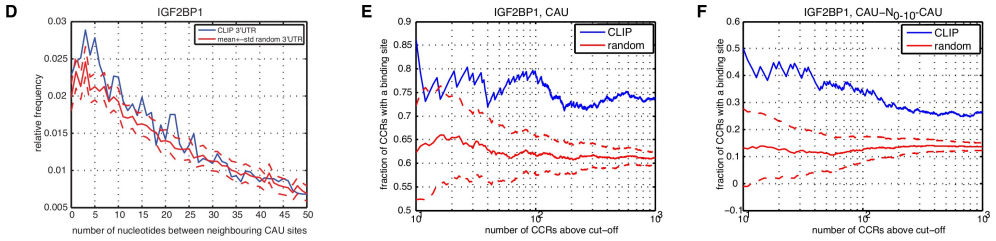


Figure 55: (D) Distance between two neighboring CAU-motifs in crosslinked IGF2BP1 PAR-CLIP clusters (blue line) and in randomized transcripts (red line). CAU-motifs are enriched within 3-5 nt distance of each other in the crosslinked regions compared to randomized sequence sets. Only IGF2BP1 is shown because IGF2BP2 and 3 show the same results. (E-F) Enrichment of the CAU (E) or CAU-N(0-10)-CAU (F) binding motif for IGF2BP1 over randomized sequence sets of the same nucleotide composition. Equivalent analyses for IGF2BP2 and IGF2BP3 yield similar results (data not shown).

2′-O-methyl oligoribonucleotides and miRNA duplexes

The following sequences were chemically synthesized on an ABI394 RNA/DNA synthesizer using 5′silyl-2′orthoester chemistry[1] (Dharmacon):

```
anti-let-7a: AACUAUACAACCUACUACCUCA-NH2;
anti-miR-10a: CACAAAUUCGGAUCUACAGGGUA-NH2;
anti-miR-15a: CGCCAAUAUUUACGUGCUGCUA;
anti-miR-15b: CACAAACCAUUAUGUGCUGCUA;
anti-miR-16: UGUAAACCAUGAUGUGCUGCUA;
anti-miR-17-5p: CUACCUGCACUGUAAGCACUUUG;
anti-miR-18a: CUAUCUGCACUAGAUGCACCUUA-NH2;
anti-miR-19a: UCAGUUUUGCAUAGAUUUGCACA;
anti-miR-19b: UCAGUUUUGCAUGGAUUUGCACA;
anti-miR-20a: CUACCUGCACUAUAAGCACUUUA;
anti-miR-20b: CUACCUGCACUAUGAGCACUUUG;
anti-miR-21: UCAACAUCAGUCUGAUAAGCUA;
anti-miR-25: UCAGACCGAGACAAGUGCAAUG;
anti-miR-27: AACUAUACAAUCUACUACCUCA;
anti-miR-30a: CUUCCAGUCGAGGAUGUUUACA-NH2;
anti-miR-30b/c: GAGUGUAGGAUGUUUACA-NH2;
anti-miR-92b: ACAGGCCGGGACAAGUGCAAUA;
anti-miR-93: CUACCUGCACGAACAGCACUUUG;
anti-miR-101: UUCAGUUAUCACAGUACUGUA;
anti-miR-103: UCAUAGCCCUGUACAAUGCUGCU;
anti-miR-106b: AUCUGCACUGUCAGCACUUUA-NH2;
anti-miR-186: AGCCCAAAAGGAGAAUUCUUUG;
anti-miR-301: GCUUUGACAAUACUAUUGCACUG;
anti-miR-378: CCUUCUGACUCCAAGUCCAGU;
miR-7/miR-7* duplex:
UGGAAGACUAGUGAUUUUGUUGU, CAACAAAUCACAGUCUGCCAUA;
miR-124/miR-124* duplex:
5â-UAAGGCACGCGGUGAAUGCCA, CGUGUUCACAGCGGACCUUGA
```

1 -NH2 indicates C6 aminolinker (Dharmacon).

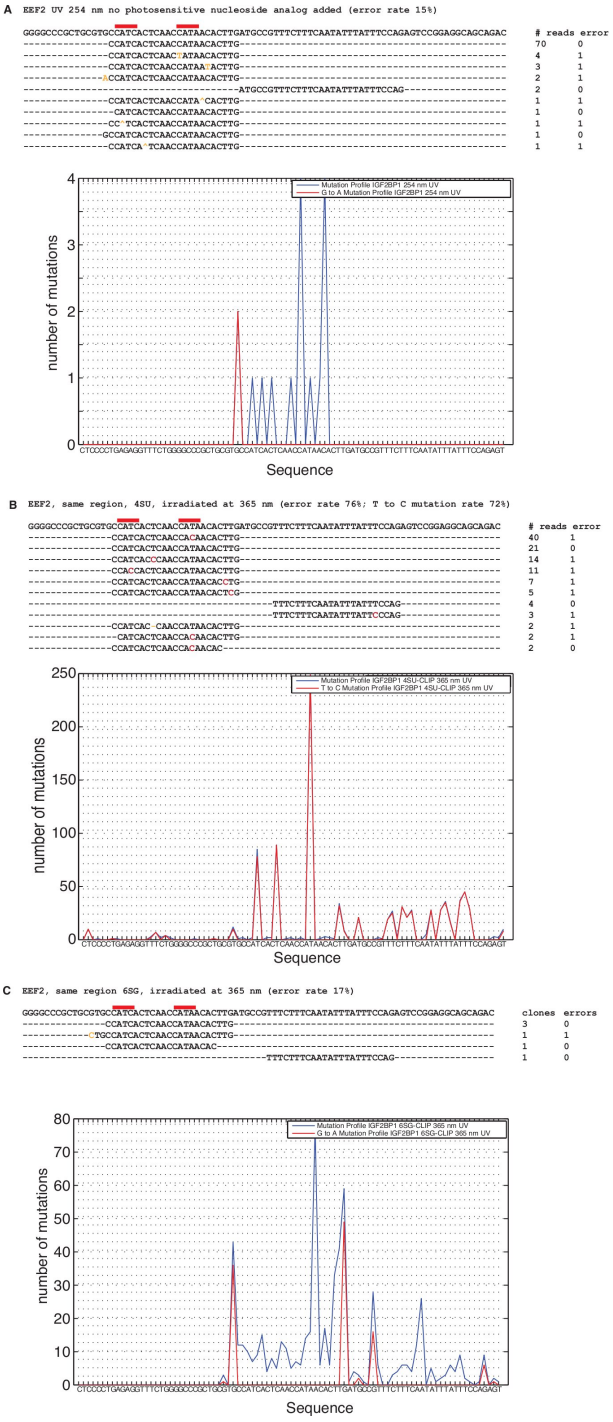


Figure 56: Comparison of a 4SU-PAR-CLIP with a 6SG-PAR-CLIP cluster and a HITS-CLIP cluster aligning to the same genomic region. Related to Figure 14. Alignment of sequences from CLIP experiments with IGF2BP1 against nucleotides 2784-2868 of the human EEF2 transcript (NM_001961). Nucleotides marked in red show the T to C changes, all other mismatches are marked in orange. Due to space limitations, not all reads that were sequenced are shown. (A) Alignment of sequences obtained from UV crosslinking at 254 nm. Lower panel: Profile for G to A mutations (red) and for any mutation (blue). (B) Alignment of sequences obtained after incorporation of 4SU into the transcript and crosslinking at 365 nm. Lower panel: mutational profile for T to C mutations (red) and for any mutation (blue). (C) Alignment of sequences obtained after incorporation of 6SG into the transcript and crosslinking at 365 nm. Lower panel: as in (A).

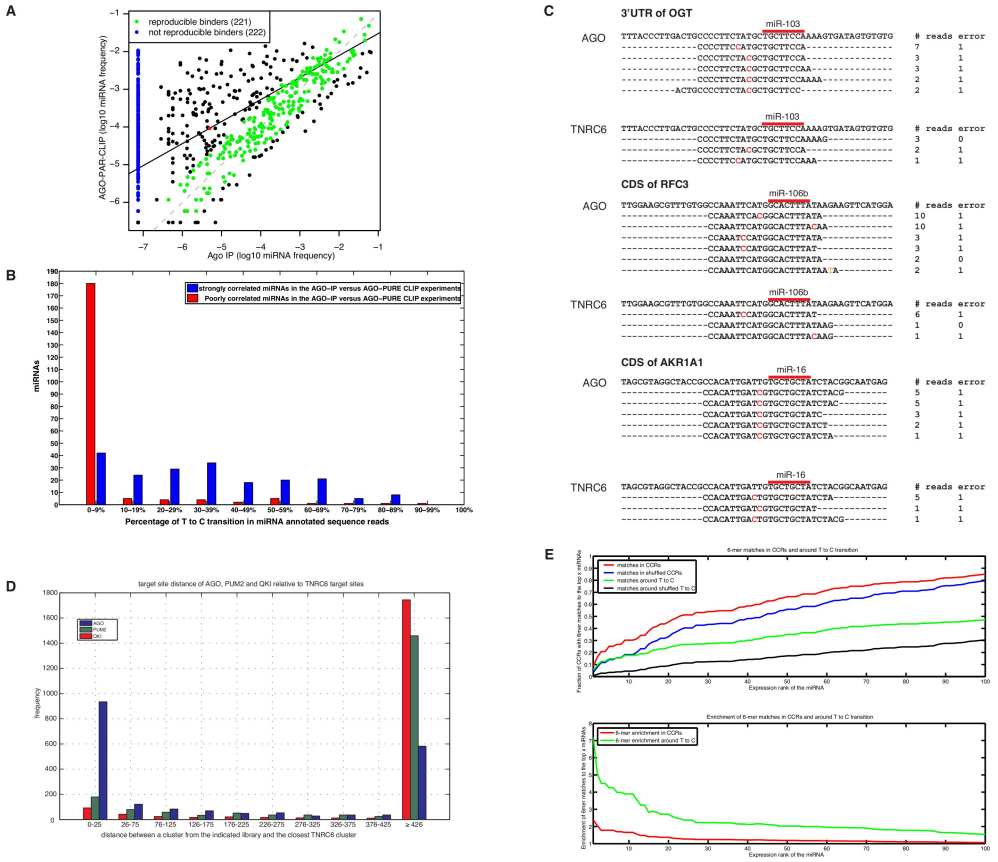


Figure 57: AGO-protein family PAR-CLIP. Related to Figure 15. (A) Principal component analysis of the relative abundance of miRNAs derived from the combination of the AGO-PAR-CLIP libraries on one hand, and the non-crosslinked AGO-IPs on the other hand. The first principal component is projected onto the plane of log10-frequency in Ago-IP vs. log10-frequency in CLIP. The slope of the principal component was 0.58. Although for many miRNAs the expression levels measured by the two methods are quite comparable, there is a subset of miRNAs whose expression in the AGO-IP is systematically lower than the expression estimated based on the AGO-PAR-CLIP data (shown in blue) (B) The miRNAs that correlate well between the AGO-IP and the AGO-PAR-CLIP data (panel A: difference in log10 frequencies in Ago CLIP vs Ago IP smaller than 0.6, shown in green) are miRNAs with high frequency of T to C mutations in the AGO-PAR-CLIP, whereas miRNAs that were sequenced at least once in the Ago CLIP but were not detected in the Ago IP (blue) have a low frequency of T to C mutations. (C)-(E) AGO and TNRC6 proteins bind to the same regions on the target transcripts. (C) Alignments of AGO PAR-CLIP and TNRC6 PAR-CLIP cDNA sequence reads to regions in the 3'UTRs of OGT (NM_181672), the CDS of RFC3 (NM_002915) and the CDS of AKR1A1 (NM_006066). Red bars indicate 8 nt seed complementary sequences and nucleotides marked in red indicate T to C mutations diagnostic of the crosslinking position. (D) The distance between TNRC6 target sites and the nearest binding sites of QKI, PUM2, AGO have been computed. The histogram shows the number of TNRC6 target sites within a given nucleotide distance from the binding site of another RNA binding protein. Approximately 950 (i.e. ca. 50%) of the CCRs from the TNRC6 PAR-CLIP experiment fall within 25 nt of a CCR from the AGO-PAR-CLIP. (E) 6-mer enrichment in the full CCRs and the region ranging from 2 nt upstream to 10 nt downstream of the predominant crosslinking site. The upper panel shows the fraction of CCRs having a 6-mer hit for the top 100 expressed miRNAs. The background set consists of dinucleotide shuffled versions of either the full CCRs or the region around the crosslinking site. The lower panel shows the enrichment of 6-mers relative to the background set in the region indicated in previous panel (full CCRs, and 13 nt around the predominant crosslinking site).

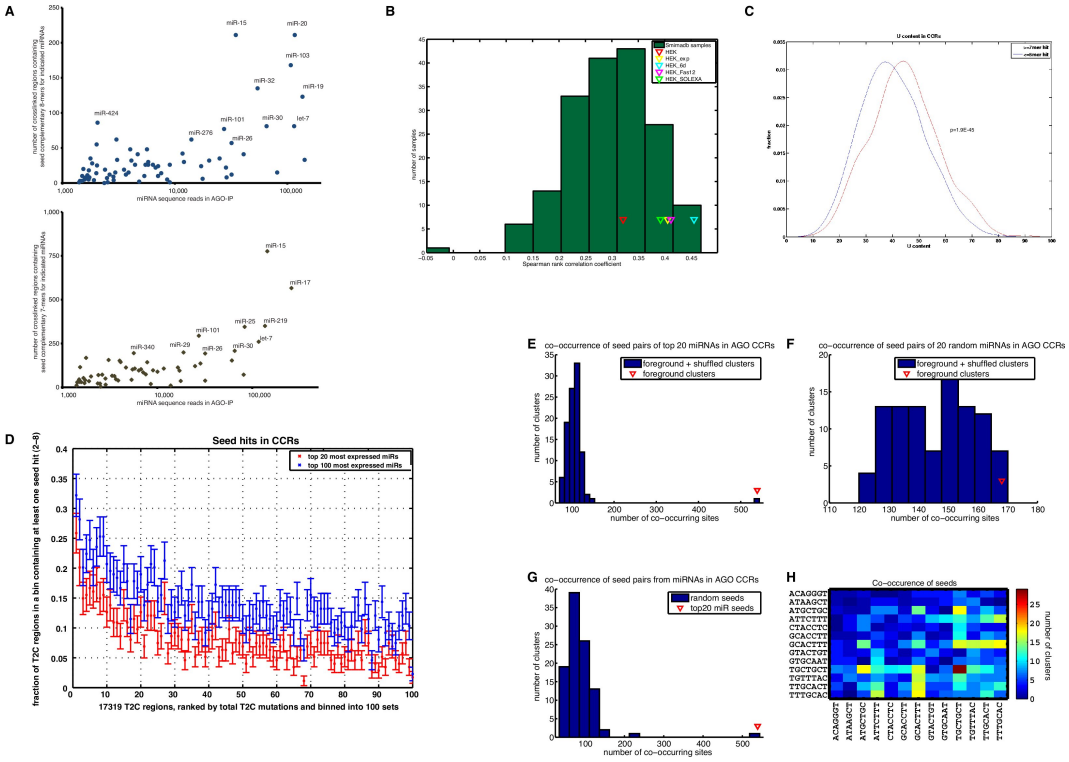


Figure 58: Seed complementary sequences from abundant HEK293 miRNAs are enriched in AGO-PAR-CLIP CCRs. Related to Figure 16. CCRs from the AGO-PAR-CLIP are enriched for target sites for the most abundant miRNAs in HEK293 cells. (A) Correlation between occurrence of 8-mer (upper panel) and 7-mer (lower panel) seed matches in the CCRs and the abundance of the corresponding miRNA seed families. (B) Spearman correlation between the number of 7-mer (2-8) seed matches in the CCRs from AGO-PAR-CLIP and the experimentally determined counts of corresponding miRNA seeds in various miRNA samples from the smiRNAdb database (www.mirz.unibas.ch/smirnadb) and the HEK293 RNA analyzed in this study. Triangles indicate different HEK293 miRNA libraries. (C) Comparison of the U content of CCRs with at least a 7-mer seed match to the top 100 most abundant miRNAs versus CCRs with at most a 6-mer seed match to the top 100 most abundant miRNAs. The mean of the distributions was significantly different (ranksum test, $p = 1.910^{-45}$). (D) The number of crosslinking events correlates with the enrichment of the CCRs in the putative binding sites for the most abundantly expressed miRNAs. The frequency of the most strongly enriched miRNA seed motif (complementary to positions 2-8 of the miRNAs) was determined in the 17,319 AGO CCRs, which were sorted by the number of U-to-C changes and grouped into bins of 100. The frequency of miRNA seed-complementary motifs in the CCRs decreases with the number of U-to-C mutations in the clusters corresponding to these CCRs. (E) Number of pairs of non-overlapping seed (pos. 2-8) matches for the 20 most abundantly expressed miRNAs in HEK 293 cells in the crosslinked regions (red triangle) and in control regions (100 sets of dinucleotide shuffled crosslinked regions). Only the experimental set shows enrichment of miRNA pairs. (F) Number of co-occurring pairs of miRNA seed matches in the AGO crosslinked regions and the shuffled control regions for 20 randomly chosen miRNAs. (G) Number of co-occurring pairs of miRNA seed matches in the AGO crosslinked regions for 100 sets of 20 randomly chosen miRNAs. (H) Heat map representation of miRNA seed match co-occurrence. Only miRNA seed matches were counted that did not overlap and could therefore be bound simultaneously by two AGO-proteins. The scale indicates the absolute number of co-occurring pairs. Matches to the seed of miR-17 co-occur with matches to the seed of miR-19/miR-130/miR-301/miR-30/miR-15/miR-16. miR-16 seed matches have the tendency to co-occur with themselves.

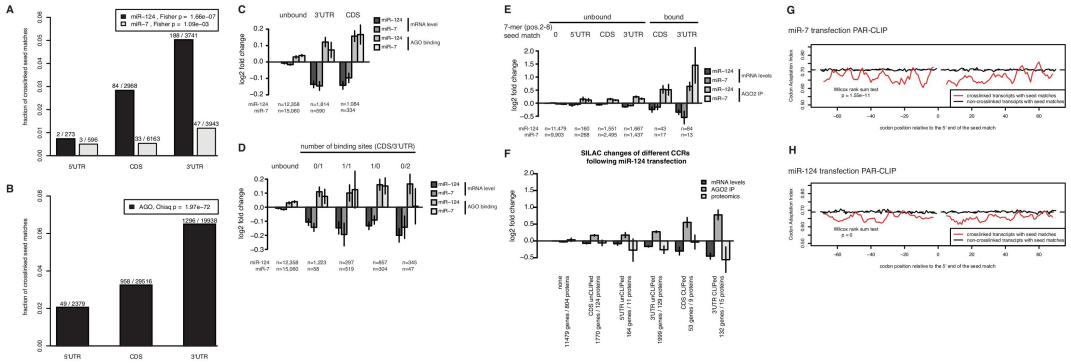


Figure 59: Properties of CCRs containing miRNA seed complementary sites. Related to Figure 17. (A) Seed complementary sequences in the 3′UTR are more efficiently crosslinked than seed complementary regions in the CDS. Fraction of crosslinked seed matches (1-7 or 2-8) for the miR-124 (dark bars) and miR-7 (light bars) transfection experiments are shown; and in (B) the fraction of crosslinked seed matches for miR-15, miR-16, miR-19, and let-7 in the ALL_AGO dataset is shown. (C) Properties of AGO-PAR-CLIP sequence read clusters obtained after miR-124 and miR-7 transfection. Transcripts with PAR-CLIP sequence read clusters identified after miR-124 and miR-7 transfection (n indicates number of transcripts considered) are bound by AGO2 and destabilized. Transcript stability (dark grey bars) was determined as in Figure 13 by comparison of mRNA-abundance of mock-transfected and miR-124 and miR-7-transfected HEK293 cells. miR-7 and miR-124 mediated AGO2 binding (light grey bars) was determined by comparing transcripts enriched by AGO2-IPs of mock transfected and miR-124 and miR-7 transfected HEK293 cells [95]. Transcripts containing PAR-CLIP sequence read clusters were categorized according to the transcript region bound by AGO2 (CDS/3′UTR). (D) Same as in (C). Transcripts were categorized in more detail according to the number and region (CDS/3′UTR) of sequence read clusters identified. (E) Same as in (C). Transcripts containing a miR-124 and miR-7 seed complementary sequence but without PAR-CLIP sequence read clusters (unbound) were compared to transcripts with PAR-CLIP sequence read clusters with miR-124 and miR-7 seed complementary sequences (bound). The unbound and bound transcripts are categorized according to regions within the transcript (5′UTR, CDS, and 3′UTR). (F) In addition to the AGO2 binding and mRNA destabilization following miR-124 transfection shown in (G) for PAR-CLIP identified transcripts, changes in protein level following miR-124 transfection (as measured by SILAC in HeLa cells by Baek et al. [8]) are indicated. (G-H) Codon adaptation index (CAI) for regions upstream and downstream of CCRs (relative to 5′ end of the seed match) found in the CDS for the (G) miR-7 and (H) miR-124 transfection experiments. The red and the black lines indicate the CAI for crosslinked and non-crosslinked transcripts, respectively.

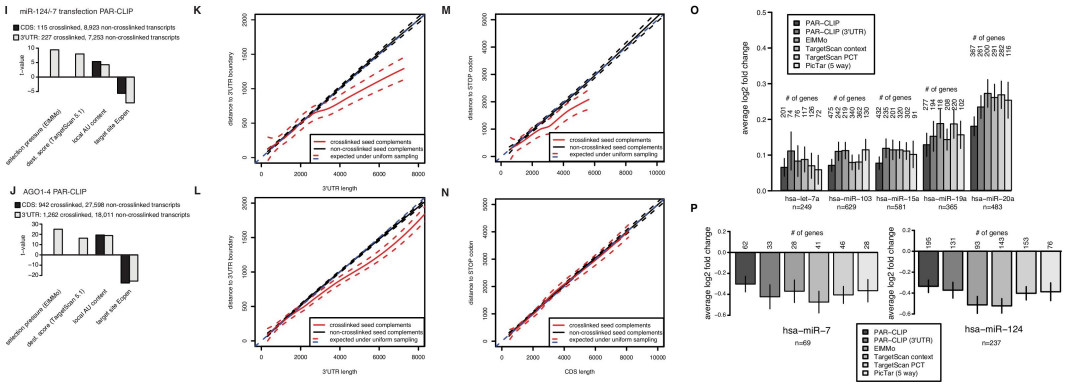


Figure 59: (I) The sequence context defines a functional miRNA binding site in the UTR as well as in the CDS. Four different criteria (selection pressure, destabilization score, local A/U content, target site openness) were compared for crosslinked transcripts containing 7-mer seed matches for a miR-124 and miR-7 and (J) the miR-15, miR-19, miR-20, and let-7 miRNA families in the AGO PAR-CLIP experiments compared to non-crosslinked transcripts containing the same 7-mer seed matches. (K) In 3'UTRs longer than 3,000 nt the crosslinked sites distribute preferentially near to the boundaries of the UTR. Distance from the region boundaries (stop codon and polyA signal, respectively) of CCRs with 7-mer seed complement regions falling in the 3'UTR to miR-124 and miR-7 in the transfection experiments (red line) and (L) 7-mer seed matches to the miR-15, miR-16, miR-19 and let-7 seed families from the AGO PAR-CLIP (red line) compared to non-crosslinked seed-matches (black lines). (M) Distance from the stop codon of CCRs falling in the CDS containing 7-mer seed matches of miR-124 and miR-7 (red line) or (N) 7-mer seed matches of the miR-15, miR-16, miR-19 and let-7 seed families (red line) compared to non-crosslinked seed-matches (black lines). Only for the miR-124 and miR-7 transfection experiments the crosslinked sites in the CDS distribute significantly closer to the stop-codon. (O) Comparison of PAR-CLIP with ElMMo, TargetScan context, TargetScan Pct, and PicTar miRNA target predictions. We determined the number of seed matches in the top 1000 CCRs for each of the indicated miRNAs. For each miRNA we selected an equal indicated number of target sites (on mRNAs found by DGE and having a signal intensity above the median on the Affymetrix mRNA microarrays) that map to the indicated number of genes, starting from those with the best score, as given by the indicated prediction method. The figure shows average log2 fold changes of mRNA targets identified by the different methods upon miRNA inhibition (of miRNAs let-7a, miR-103, miR-15a, miR-19a, miR-20). (P) Average log2 fold changes of mRNA targets identified by various methods upon miR-7 and miR-124 transfection.

Plasmids

Plasmids pENTR4 IGF2BP1-3, QKI, AGO1-4, TNRC6A-C and PUM2 were generated by PCR amplification of the respective coding sequences (CDS) followed by restriction digest with SalI and NotI and ligation into pENTR4 (Invitrogen). pENTR4 IGF2BP1,-2, and -3 were recombined into pFRT/TO/FLAG/HA-DEST destination vector (Invitrogen) using GATEWAY LR recombinase (Invitrogen) according to manufacturer's protocol to allow for doxycycline-inducible expression of stably transfected FLAG/HA-tagged protein in Flp-In T-REx HEK293 cells (Invitrogen) from the TO/CMV promoter. pENTR4 QKI and pENTR4 PUM2 were recombined into pFRT/FLAG/HA-DEST for constitutive expression in Flp-In T-REx HEK293 cells.

Plasmids for bacterial expression of N-terminally His6-tagged IGF2BP1, 2, and 3 in E. coli were generated by ligation of CDS into pET16 (Novagen). The plasmid for bacterial expression of N-terminally His6-tagged QKI was generated by LR recombination of pENTR4 QKI with pDEST17 (Invitrogen). The plasmids described in this study can be obtained from Addgene (www.addgene.org).

Antibodies

Polyclonal rabbit antibodies against IGF2BP1, 2, and 3 were generated by injection of synthetic peptides corresponding to amino acids 561-573, 264-275, and 567-579, respectively. Rabbit anti-QKI (BL1040) was purchased from Bethyl Laboratories.

Recombinant protein expression and purification

pET16 IGF2BP1,-2, and -3 and pDEST17-QKI plasmids, encoding an N-terminal His6-tag, were transformed in E. coli STAR(DE3) (Invitrogen). Cells were grown in LB medium supplemented with 50 µg/ml ampicillin at 37°C to $A_{600} = 0.6$. The cells were cooled to 25°C, protein synthesis was induced by addition of IPTG to a final concentration of 1 mM, cells were harvested 3 h later. The cell pellet was resuspended in 10 ml lysis buffer (50 mM Tris-HCl pH 8.0, 300 mM KCl, 5 mM $MgCl_2$, 0.1% Triton X-100, and complete EDTA-free protease inhibitor (Roche)) per gram cell pellet. All the following steps were carried out at 4°C. Cells were resuspended in lysis buffer and incubated with 1 mg/ml lysozyme for 30 min and sonicated to reduce viscosity. Insoluble material was removed by centrifugation at 12,000xg for 20 min. For His-tag affinity selection, the supernatant was incubated with 250 µl HIS-Select Cobalt Affinity Gel (Sigma) per 10 ml cell supernatant for 1 h. The gel was washed three times with 10 gel volumes of wash buffer (50 mM Tris-HCl, pH 8.0, 300 mM Kcl, 5 mM $MgCl_2$, 1 mM DTT, 0.1 % Triton X-100, 25 mM imidazol, and complete EDTA-free protease inhibitor (Roche)). His-tagged proteins were eluted in 3 gel volumes of elution buffer (50 mM Tris-HCl pH 8.0, 300 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 0.1% Triton X-100, 250 mM imidazol, and complete EDTA-free protease inhibitor (Roche)). The eluted proteins were applied to a Heparin column equilibrated in 20 mM Tris-HCl pH 7.8, 5 mM $MgCl_2$, 100 mM KCl, 1 mM DTT, 0.1% Triton X-100, 10% glycerol. Proteins were eluted with a KCl gradient (0.5 - 1.5 M) in 20 mM Tris-HCl, pH 7.8, 5 mM $MgCl_2$, 1 mM DTT, 0.1% Triton X-100, 10% glycerol. His6-

IGF2BP1, -2, and -3 eluted at 550 to 650 mM KCl and His6-QKI at 1.1 M KCl.

Electrophoretic mobility-shift analysis

Radiolabeled RNA (100 pM) was incubated with recombinant His6-IGF2BP2 protein at indicated concentrations and 100 ng tRNA in binding buffer (20 µl of 20 mM Tris-HCl, pH 7.8, 140 mM KCl, 2 mM $MgCl_2$ and 0.1% Triton X-100 at 30°C) for 1 h. After addition of 6 µl loading dye (40% glycerol, bromophenol blue in binding buffer), the reaction mixture was loaded onto a native 6% acrylamide gel containing 0.5x TBE, running at 200 V for 1 h at room temperature, using 0.5x TBE as running buffer. Radiolabeled RNA (1 nM) was incubated with recombinant His6-QKI protein at various concentrations and 100 ng tRNA in 20 µl of binding buffer (20 mM HEPES-KOH, pH 7.4, 330 mM KCl, 10 mM MgCl2, 0.1 mM EDTA and 0.01% IGEPAL CA630 (Sigma)). After addition of 6 µl loading dye (40% glycerol, bromophenol blue in binding buffer), the solution was loaded onto a native 10% acrylamide gel containing 0.5x TBE, running at 200 V for 2 h at room temperature, using 0.5x TBE as running buffer. The protein-bound RNA and the free RNA were quantified using a phosphorimager.

Cell lines and culture conditions

HEK293 T-REx Flp-In cells (Invitrogen) were grown in D-MEM high glucose with 10% (v/v) fetal bovine serum, 1% (v/v) 2 mM L-glutamine, 1% (v/v) 10,000 U/ml penicillin/10,000 µg/ml streptomycin, 100 µg/ml zeocin and 15 µg/ml blasticidin. Cell lines stably expressing FLAG/HA-tagged proteins were generated by co-transfection of pFRT/TO/FLAG/HA or pFRT/FLAG/HA constructs with pOG44 (Invitrogen). Cells were selected by exchanging zeocin with 100 µg/ml hygromycin. Expression of FLAG/HA-IGF2BP1, -2, -3 and TNRC6A, B and C was induced by addition of 250 ng/ml doxycycline 15 to 20 h before crosslinking.

miRNA profiling

miRNAs were extracted from FLAG/HA-AGO immunoprecipitates as described in Meister et al. [153]. miRNAs from immunoprecipitates and the lysate were cloned and Solexa-sequenced [89] using following bar-coded 5' adapters:

```
AGO1-IP: TCTAGTCGTATGCCGTCTTCTGCTTGT
AGO2-IP: TCTCCTCGTATGCCGTCTTCTGCTTGT
AGO2-IP: TCTGATCGTATGCCGTCTTCTGCTTGT
AGO3-IP: TTAAGTCGTATGCCGTCTTCTGCTTGT
Lysate: TCACTTCGTATGCCGTCTTCTGCTTGT
```

Determination of incorporation levels of 4-thiouridine into total RNA

Flp-In HEK293 were grown in medium supplemented with 100 µM 4SU 16 h prior to harvest. As a control, cells grown without 4SU addition were also harvested. 3 volumes of Trizol reagent (Sigma) were added to the washed cell pellets and total RNA was extracted according to manufactures instructions. Total RNA was further puri-

fied using Qiagen RNAeasy according to the manufacturer's protocol. To prevent oxidization of 4SU during RNA isolation and analysis, 0.1 mM dithiothreitol (DTT) was added to the wash buffers and subsequent enzymatic steps. Total RNA was digested and dephosphorylated to single nucleosides for HPLC analysis [5]. Briefly, in a 30 µl volume, 40 µg of purified total RNA were incubated for 16 h at 37°C with 0.4 U bacterial alkaline phosphatase (Worthington Biochemical) and 0.09 U snake venom phosphodiesterase (Worthington Biochemical). As a reference standard, synthetic 4SU-labeled RNA, CGUACGCGGAAUACUUCGA(4SU)U was used and also subjected to complete enzymatic digestion. The resulting mixtures of ribonucleosides were separated by HPLC on a Supelco Discovery C18 (bonded phase silica 5 µM particle, 250 x 4.6 mm) reverse phase column (Bellefonte PA, USA). HPLC buffers were 0.1 M TEAA in 3% acetonitrile (A) and 90% acetonitrile in water (B). The gradient was isocratic 0% B for 15 min, 0 to 10 % B for 20 min, 10 to 100% B for 30 min, and a 5 min 100% B wash applied between runs to clean the HPLC column.

UV 254 nm and UV 365 nm crosslinking

For UV crosslinking, cells were washed once with ice-cold PBS while still attached to the plates. PBS was removed completely and cells were irradiated on ice with 254 nm UV light (0.15 J/cm^2), or 365 nm UV light for cells treated for 14 h with 100 µM nucleoside analogs (0.15 J/cm^2) in a Stratalinker 2400 (Stratagene), equipped with light bulbs for the appropriate wavelength. Cells were scraped off with a rubber policeman in 1 ml PBS per plate and collected by centrifugation at 500xg for 5 min.

Cell lysis and first partial RNase T1 digestion

The pellets of cells crosslinked with UV 365 nm were resuspended in 3 cell pellet volumes of NP40 lysis buffer (50 mM HEPES, pH 7.5, 150 mM KCl, 2 mM EDTA, 1 mM NaF, 0.5% (v/v) NP40, 0.5 mM DTT, complete EDTA-free protease inhibitor cocktail (Roche)) and incubated on ice for 10 min. The typical scale of such an experiment was 3 ml of cell pellet. The cell lysate was cleared by centrifugation at 13,000xg. RNase T1 (Fermentas) was added to the cleared cell lysates to a final concentration of 1 U/µl and the reaction mixture was incubated in a water bath at 22°C for 15 min and subsequently cooled for 5 min on ice before addition of antibody-conjugated magnetic beads.

Immunoprecipitation and recovery of crosslinked target RNA fragments

PREPARATION OF MAGNETIC BEADS 10 µl of Dynabeads Protein G magnetic particles (Invitrogen) per ml cell lysate were washed twice with 1 ml of citrate-phosphate buffer (4.7 g/l citric acid, 9.2 g/l Na2HPO4, pH 5.0) and resuspended in twice the volume of citrate-phosphate buffer relative to the original volume of bead suspension. 0.25 µg of anti-FLAG M2 monoclonal antibody (Sigma) per ml suspension was added and incubated at room temperature for 40 min. Beads were then washed twice with 1 ml of citrate-phosphate buffer to remove un-

bound antibody and resuspended again in twice the volume of citrate-phosphate buffer relative to the original volume of bead suspension.

IMMUNOPRECIPITATION (IP), SECOND RNASE T1 DIGESTION, AND DEPHOSPHORYLATION 10 µl of freshly prepared antibody-conjugated magnetic beads per ml of partial RNase T1 treated cell lysate were added and incubated in 15 ml centrifugation tubes on a rotating wheel for 1 h at 4°C. Magnetic beads were collected on a magnetic particle collector (Invitrogen). Manipulations of the following steps were carried out in 1.5 ml microfuge tubes. The supernatant was removed from the bead-bound material. Beads were washed 3 times with 1 ml of IP wash buffer (50 mM HEPES-KOH, pH 7.5, 300 mM KCl, 0.05% (v/v) NP40, 0.5 mM DTT, complete EDTA-free protease inhibitor cocktail (Roche)) and resuspended in one volume of IP wash buffer. RNase T1 (Fermentas) was added to obtain a final concentration of 100 U/µl, and the bead suspension was incubated in a water bath at 22°C for 15 min, and subsequently cooled for 5 min on ice. Beads were washed 3 times with 1 ml of high-salt wash buffer (50 mM HEPES-KOH, pH 7.5, 500 mM KCl, 0.05% (v/v) NP40, 0.5 mM DTT, complete EDTA-free protease inhibitor cocktail (Roche)) and resuspended in one volume of dephosphorylation buffer (50 mM Tris-HCl, pH 7.9, 100 mM NaCl, 10 mM MgCl2, 1 mM DTT). Calf intestinal alkaline phosphatase (NEB) was added to obtain a final concentration of 0.5 U/µl, and the suspension was incubated for 10 min at 37°C. Beads were washed twice with 1 ml of phosphatase wash buffer (50 mM Tris-HCl, pH 7.5, 20 mM EGTA, 0.5% (v/v) NP40) and twice with 1 ml of polynucleotide kinase (PNK) Buffer (50 mM Tris-HCl, pH 7.5, 50 mM NaCl, 10 mM MgCl2, 5 mM DTT). Beads were resuspended in one original bead volume of PNK buffer.

RADIOLABELING OF RNA SEGMENTS CROSSLINKED TO IMMUNOPRECIPITATED PROTEINS To the bead suspension described above, γ-32P-ATP was added to a final concentration of 0.5 µCi/µl and T4 PNK (NEB) to 1 U/µl in one original bead volume. The suspension was incubated for 30 min at 37°C. Thereafter, non-radioactive ATP was added to obtain a final concentration of 100 µM and the incubation was continued for another 5 min at 37°C. The magnetic beads were then washed 5 times with 800 µl of PNK Buffer and resuspended in 70 µl of SDS-PAGE Loading Buffer (10% glycerol (v/v), 50 mM Tris-HCl, pH 6.8, 2 mM EDTA, 2% SDS (w/v), 100 mM DTT, 0.1% bromophenol blue).

SDS-PAGE AND ELECTROELUTION OF CROSSLINKED RNA-PROTEIN COMPLEXES FROM GEL SLICES The radiolabeled bead suspension was incubated for 5 min at 95°C and vortexed. The magnetic beads were separated on a magnetic separator and 40 µl of supernatant were loaded per well of an SDS-PAGE. The gel was analyzed by phosphorimaging. The radioactive RNA-protein complex migrating at the expected molecular weight of the target protein was excised from the gel and electroeluted in a D-Tube Dialyzer Midi (Novagen) in 800 µl SDS running buffer according to the instructions of the manufacturer.

PROTEINASE K DIGESTION An equal volume of 2x Proteinase K Buffer (100 mM Tris-HCl, pH 7.5, 150 mM NaCl, 12.5 mM EDTA, 2%

(w/v) SDS) with respect to the electroeluate was added, followed by the addition of Proteinase K (Roche) to a final concentration of 1.2 mg/ml, and incubation for 30 min at 55°C. The RNA was recovered by acidic phenol/chloroform extraction followed by a chloroform extraction and an ethanol precipitation. The pellet was dissolved in 10.5 µl water.

cDNA library preparation and deep sequencing

The recovered RNA was carried through a cDNA library preparation protocol originally described for cloning of small regulatory RNAs [89]. The first step, 3′ adapter ligation, was carried out as described on a 20 µl scale using 10.5 µl of the recovered RNA. UV 254 nm crosslinked RNAs were processed using standard adapter sets, followed by PCR to introduce primers compatible with 454 sequencing; UV 365 nm crosslinked sample RNAs were processed using Solexa sequencing adapter sets. Depending on the amount of RNA recovered, 5′-adapter-3′-adapter products without inserts may be detected after amplification of the cDNA as additional PCR bands. In such case, the longer PCR product of expected size was excised from a 3% NuSieve low-melting point agarose gel, eluted from the gel pieces with the Illustra GFX-PCR purification kit (GE Healthcare) and Solexa sequenced.

Oligonucleotide transfection and mRNA array analysis

siRNA, miRNA and 2′-O-methyl oligonucleotide transfections of HEK293 T-REx Flp-In cells were performed in 6-well format using Lipofectamine RNAiMAX (Invitrogen) as described by the manufacturer. Total RNA of transfected cells was extracted using TRIZOL following the instructions of the manufacturer. The RNA was further purified using the RNeasy purification kit (Qiagen). 2 µg of purified total RNA was used in the One-Cycle Eukaryotic Target Labeling Assay (Affymetrix) according to manufacturer's protocol. Biotinylated cRNA targets were cleaned up, fragmented, and hybridized to Human Genome U133 Plus 2.0 Array (Affymetrix). For details of the analysis, see Bioinformatics section.

Generation of Digital Gene Expression (DGEX) libraries

1 µg each of total RNA from HEK293 cells inducibly expressing tagged IGF2BP1 before and after induction was converted into cDNA libraries for expression profiling by sequencing using the DpnII DGE kit (Illumina) according to instructions of the manufacturer. For details of the analysis, see Bioinformatics section.

C.4 BIOINFORMATICS ANALYSES

Adapter removal and sequence annotation

The basic method for removing adaptors and assigning a functional annotation to the sequence reads was described in Berninger et al. [16]. Briefly, we used an in-house ends-free local alignment algorithm (score parameters: 2 for match, -3 for mismatch, -2 for gap opening, -3 for

gap extension) to align the Solexa adapter to the 3′ end of each sequence read, allowing for the possibility that the adapter was not completely sequenced[2]. We removed from the reads the fragments that aligned to the adaptor as long as the number of matches exceeded that of mismatches by at least 3. Sequences that were either too short (less than 20 nt) or too repetitive (using a cut-off of 0.7 and 1.5 in the entropy of the mono- and dinucleotide distributions, respectively, of individual sequence reads [16]) were discarded because they would probably map to multiple genomic locations. The remaining sequences were mapped to the hg18 version of the human genome assembly that was downloaded from the University of California at Santa Cruz (http://genome.cse.ucsc.edu) and to a database of sequences whose function (rRNA, tRNA, sn/snoRNA, miRNA, mRNA, etc.) is already known. These were obtained from the sources specified in Berninger et al. [16]. The Oligomap algorithm [16] was used for this purpose, and all the perfect and 1-error (mismatch or insertion or deletion (indel) mappings were obtained. Based on the GMAP [232] genome mapping of human mRNA transcripts from NCBI downloaded on November 4th, 2008, we determined whether the sequence reads mapped to intronic or exonic regions of genes. Based on the coding region annotation of transcripts in GenBank, we determined whether the exonic sequence reads originated from the 5′UTR, CDS or 3′UTR.

Generation of clusters of mapped sequence reads

For subsequent analyses only sequence reads that were at least 20 nucleotides long and mapped uniquely to the genome with at most one error were used. A single-linkage clustering of the sequence reads was performed, with two reads being placed in the same cluster if they overlapped by at least one nucleotide in their genomic mappings. Each cluster was then annotated based on the functional annotation of sequence reads that covered most of the cluster length. We then considered all the mRNA-annotated clusters containing at least 5 mRNA-annotated sequence reads, and we defined a scoring scheme to identify the clusters that had the highest probability of being real crosslinking sites (see below: Identification of high confidence clusters).

Analysis of the mutational spectra

From the clusters defined above, all sequence reads were used that mapped uniquely and with one error (mismatch or indel) to the genome to infer the mutational bias of the method. For each library, we calculated the proportion of mutations involving each of the four nucleotides as well as the proportion of each of the four nucleotides in the crosslinked sequence reads (see Figure 53B,C).

Identification of high-confidence clusters

We used the crosslinked clusters of PUM2 and QKI, to define criteria for selecting high-confidence binding sites. The criteria that we tested reflected the mechanistic aspects of generating the sequence reads. Our

2 Software can be downloaded from http://www.mirz.unibas.ch/restricted/clipdata/RESULTS/index.html

preliminary analysis revealed that T to C mutations are by far the most frequently observed mutations in these data sets, and that they are most frequent inside or in the immediate vicinity of the binding motifs as opposed to the rest of the sequence (see Figures 12E, 13E, and 14E). This suggested that the observed mutational bias is directly linked to the crosslinking event and should thus be a good criterion for separating true crosslinked sites from background sequence reads. The preliminary analysis also indicated a strong bias for having G nucleotides at the last position of a sequence read and also at the genomic position immediately upstream of a sequence read. This bias reflects the sequence specificity of the RNase T1, and may again help in the identification of sequence reads that map to multiple sites or for discriminating random RNA turnover products unrelated to RNase T1 treatment. Finally, we observed that many clusters with abundantly sequenced reads contained more than one position with a T to C mutation. The results of testing these criteria for their ability to select clusters that contained the known binding motif for QKI and PUM2 are shown in Figure 54. For QKI, binding motifs were defined as occurrences of ACUAA or AUUAA, which we identified from a very small number of clusters. The first of these motifs was also identified previously through SELEX experiments [71]. For PUM2, in order to account for additional motif variants besides the consensus UGUANAUA, binding motifs were identified as matches to the weight matrix (as inferred by MotEvo [218] that resulted from the motif search (see below). We found that ranking of the clusters by the number of T to C mutations in all reads in the clusters of sequence reads leads to the strongest enrichment in clusters with a binding site (Figure 54). The figures show the fraction of the crosslinked clusters that contain at least one occurrence of the known binding motif as a function of the number of clusters that passed a given cut-off in the selection criterion (e.g. total number of sequence reads, total number of T to C mutations, total number of sequence reads with a G at position -1 relative to their genomic locus). The cut-off decreases from the left to the right of the x-axis. It is clear that, particularly for PUM2, the number of T to C mutations strongly correlates with the presence/absence of the motif in the cluster. For comparison, we also show the same plots when using as the ranking criterion not the total number of T to C mutations in the cluster, but just the total number of sequence reads per cluster. For QKI, this leads to a significantly lower enrichment of clusters with recognition elements. We also investigated how the fraction of clusters with the known binding motif depends on the number of distinct crosslinking positions (i.e. positions with at least one T to C mutation) inside the cluster (Figure 54). The fraction of clusters with a binding site increases steadily from 0 to 5 crosslinking positions for both proteins, with the strongest increase from 0 to 1 for PUM2 and between 0 and 2 crosslinking positions for QKI. When requiring that at least two positions with T to C mutations are present in the cluster, the fraction of clusters with a binding site increases roughly by 20 % for PUM2, and by more than 40 % for QKI. These considerations led us to the following procedure for defining high confidence clusters for any given RBP. We first selected all the clusters with at least two crosslinking positions and, secondly, within this subset, we ranked all clusters by the total number of T to C mutations in all sequence reads in the cluster.

Extraction of peaks and crosslink-centered regions (CCRs) from sequence read clusters

From each ranked, mRNA-annotated cluster, a peak region, defined as a 32-nt long region with the highest average sequence read density, was extracted. Because the T to C mutation was diagnostic for the site of crosslinking, we focused our motif analysis on regions anchored at the position in a cluster with the most T to C mutations. We then investigated the mutational profile around this position and we found that this profile approaches the background profile after about 20 nt to the left and right of the main site of T to C mutations. Thus, these 41-nt long regions centered on the main site of T to C mutations are most likely to contain the binding sites and we focused our motif search on these regions.

RNA recognition element search

For the motif search defining the core of a RNA recognition site we selected, for each RBP, the top 100 high confidence clusters, defined as described above. We selected the 41-nt region centered on the main T to C mutation site and searched for over-represented sequence motifs using PhyloGibbs [197]. We used a first-order Markov model as the background model and searched each set of sequences for three motifs of lengths varying between 4 and 8 nt, demanding an expected total number of 50 motifs. For each parameter setting, we performed five replicate runs. This generally resulted for each RBP in various shifted versions of the same motif. Therefore we hierarchically clustered all the weight matrices that we obtained from these runs, allowing for partial overlap of at least 4 nucleotides between pairs of weight matrices. In the clustering procedure, two weight matrices were fused if the posterior probability of their stemming from the same as opposed to two different probability distribution was larger than 0.2 (for a description of the Bayesian calculation, see Berninger et al. [16, section 4.1]). Replicating this procedure multiple times yielded very similar results (not shown). For each protein, we selected the largest cluster of weight matrices, i.e. the cluster that contained most of the weight matrices that we obtained in replicate runs, and created the final weight matrix by summing up the counts for each nucleotide of the weight matrices belonging to this cluster. Since the clustering procedure also allows the fusion of only partially overlapping weight matrices, the resulting weight matrices are typically longer (roughly 10 nucleotides) than the motif length that we imposed in individual runs, and can contain stretches of low information content. We therefore selected for each RBP, the window with highest information content. For PUM2 and QKI, the length of this window was 8 and 6 nt, respectively, in accordance with the known or expected consensus motifs [71, 76], respectively. For the IGF2BPs, we chose a window length of 4 nt, which is believed to be the size of binding motifs of KH-domains [217]. To identify binding sites in PUM2 clusters of aligned sequence reads using the inferred weight matrix, we used the MotEvo algorithm [218], which is based on a hidden Markov model that models the input sequences as contiguous stretches of nucleotides drawn from a background or a weight matrix model. We chose for the background a first order Markov model (which makes every nucleotide dependent

on the preceding nucleotide in the sequence). The background model parameters (dinucleotide frequencies) were estimated from the set of input sequences. MotEvo was run in the prior-update mode, meaning that we attempted to find the prior probabilities for sites and background that maximize the likelihood of the sequence data. MotEvo generates as an output a list of sites for the given input weight matrix as well as their corresponding posterior probabilities. Note that not all matches to the weight matrix are reported, but only the subset of matches whose corresponding sequence is more likely under the weight matrix model than the background model. We chose a cut-off of 0.4 on the posterior probability to define the set of binding sites.

Determination of the location of sequence read clusters within functional mRNA regions

For each RBP, we investigated whether clusters of mapped sequence reads preferentially originated in 5'UTR, CDS or 3'UTR (Figure 53A). As a result of our annotation pipeline, we could assign probabilities to each cluster to belong to either 5'UTR, CDS and 3'UTR based on the annotation of individual sequence reads within the cluster (see above). Taking together these probabilities for all clusters, we obtained estimates of the numbers of clusters originating in each of these three regions. We compare these numbers to those that we would expect if clusters were sampled uniformly from anywhere along the transcripts. This would for instance result in many more clusters from 3' compared to 5'UTR regions simply because 3'UTRs tend to be longer than the 5'UTRs. We determined all the transcripts to which a cluster mapped, and based on the GenBank annotation of the CDS of these transcripts, we calculated the fraction of the cluster nucleotides that fell in the 5'UTR (f_5), CDS (f_{CDS}), and 3'UTR (f_3). In the cases in which the cluster mapped to several transcripts belonging to the same gene, these fractions were averaged over all transcripts. The expected proportion of nucleotides sequenced from each region was calculated by summing these fractions for all clusters. The variance was determined by noting that the probability that a nucleotide was sampled from a particular region, e.g. 5'UTR, is Bernoulli distributed with parameter f_5, which has a variance of $f_5(1 - f_5)$. The total variance was then computed as the sum of all the variances.

Distance distribution between consecutive CAU-motifs in the IGF2BP RNA binding sites

Since each of the IGF2BPs has 4 KH domains and we found only one clear motif, we hypothesized that all KH domains have the same or a very similar binding specificity. In analogy to what has been observed for the neuronal RBP involved in splicing, Nova [215], we propose that the binding specificity of the IGF2BPs arises from the concerted action of several KH-domains that each recognize the same 4 letter sequence (CAUH), which should be apparent by a preferred spacing between subsequent occurrences of the motif as determined by the distance of corresponding KH-domains in the structure of the IGF2BPs. We calculated, for each IGF2BP separately, the distribution of distances between subsequent occurrences of the CAU-motif in clusters unambiguously derived from the 3'UTR of protein coding genes. We restricted our-

selves to these clusters since 3'UTR regions are overrepresented in clusters of the IGF2BPs and each region, 5'UTR, CDS and 3'UTR, has different sequence biases that need to be taken into account when modeling background distributions. In order to reduce boundary effects due to the finite length of the clusters, we extended each cluster region 32 nt to the right and left[3]. We then compared this distance distribution to the distance distribution of consecutive occurrences of the CAU motif in randomly chosen 3'UTR regions of the same length distribution as the clusters of mapped sequence reads. To estimate the mean and standard deviation of the relative frequency of each inter-motif distance in the background dataset, we repeated the random selection of 3'UTR regions 1000 times.

Enrichment of identified binding motifs in all clusters

We defined the binding motifs for PUM2, QKI and IGF2BPs using a subset of high-confidence clusters for each protein. To determine to what extent these motifs were indeed representing the binding sites of the proteins in the complete data sets, we collected, for each protein and for each cluster, all the respective crosslink-centered regions (CCRs) and ranked them by the number of T to C mutations. We then calculated for varying cut-offs on the number of T to C mutations the fraction of clusters above the given cut-off that contain at least one binding site (Figure 55, blue traces). The binding site was defined to be UGUANAUA for PUM2, ACUAA or AUUAA for QKI and CAU or two CAUs separated by no more than 10 nucleotides for the IGF2BPs. To estimate the number of sites expected by chance, we generated 1000 sets of random sequences with the same nucleotide frequencies as the CCRs (dinucleotide shuffling for PUM2 as well as QKI and mononucleotide shuffling for the IGF2BPs, due to the small length of the binding motif). For all proteins, the CCRs are clearly enriched in the respective binding motifs. The enrichment is strongest for PUM2, which has the longest recognition motif. For the IGF2BPs, the enrichment for the CAU-spacer-CAU motif is much stronger than for the CAU motif due to the clustering of the CAU motif (see previous section). For PUM2, we additionally determined the enrichment only for the first half of motif UGUA. This short motif is clearly enriched and is contained in more than 72 percent of all CCRs, suggesting the presence of other variants of the PUM2 motif besides the consensus UGUANAUA.

Analysis of siRNA knockdown experiments

We imported the CEL files into the R software (`http://www.R-project.org`) using the BioConductor affy package [74]. The transcript probe set intensities were background-corrected, adjusted for non-specific binding and quantile normalized with the GCRMA algorithm [233]. Probe sets with more than 6 of the 11 probes mapping ambiguously to the genome were discarded, as were probe sets that mapped to multiple genes. We then collected all probe sets matching a given gene, and we selected for further analysis the RefSeq transcript with median 3'UTR length corresponding to that gene. In total 16,063 transcripts were identified. The log-intensity of probe sets mapping to the gene

3 The genomic regions are shown on the website `http://www.mirz.unibas.ch/restricted/clipdata/RESULTS/index.html`

were then averaged to obtain the expression level per RefSeq transcript. The changes of transcript abundances were computed as the logarithm of the ratio of transcript expression in the cocktails of siRNA treated samples and mock-transfected cells.

To study the effect of individual proteins on the mRNA stability of their targets, we performed the following analysis. We first made the links between clusters of mapped Solexa sequence reads and expression data based on the NCBI Gene ID. That is, both the transcripts that were crosslinked and those whose expression was measured on microarrays have associated Gene IDs in the Gene database of NCBI. We mapped both the mapped sequence read clusters as well as the transcripts on microarrays to their corresponding genes, and thus identified which genes that were represented on microarrays have been crosslinked. From this set of genes we removed those that are likely off-targets of the transfected siRNAs. As previous studies showed, complementarity to the first 8 nucleotides of the miRNA is a good indicator that the transcript will be downregulated by a miRNA or siRNA, so we defined as putative off-targets those genes whose representative RefSeq transcripts carried such complementary sites in their 3'UTR. We divided the list of genes sorted by the maximum score of any cluster associated with a given gene. In order to improve the target identification and the assessment of the target response, we used some specific information that was available for individual data sets. For instance, for the IGF2BPs we only considered clusters with at least 2 positions of T to C changes, because we previously observed that this criterion improves the accuracy of target identification for the PUM2 and QKI. Thus, for the IGF2BPs we divided the bound transcripts into the following bins, top 100 genes, 101th - 300th genes, 301th -500th genes and 501th -1000th genes, 1001th-2000th, 2001th-3497th, and calculated the log2fold change of transcript abundance. To determine whether the siRNA knockdown has an effect on mRNA stability, we compared these distributions with the distribution of log-fold changes of genes that did not have any associated clusters from CLIP analysis. For QKI, we performed the same analysis starting from clusters with a single T to C mutation site, but that additionally contained the QKI motif.

Generation and ranking of clusters of mapped sequence reads for AGO and TNRC6 family PAR-CLIP

To generate sequence read clusters for the cDNA libraries from the AGO and TNRC6 PAR-CLIP we used sequence reads of at least 20 nt in length and with unique, perfect or 1-error mapping to the genome. We clustered the reads with single-linkage criterion, meaning that we placed two reads in the same cluster if they overlapped by at least one nucleotide in their genomic mappings. We then selected the clusters that contained at least 5 mRNA-annotated reads and at least 2 positions at which T to C mutations occurred in the sequence reads relative to the genomic sequence, and we ranked them by the total number of T to C mutations which, as we described above, is indicative of the number of crosslinks.

Definition of CCRs for sequence read clusters of AGO and TNRC6 PAR-CLIP

In each ranked, mRNA-annotated cluster we identified the position with the largest number of T to C mutations, and we constructed the mutation frequency profile around this position. We found that this profile approaches the background after about 20 nucleotides to the left and right of the position with the maximum number of T to C changes, and we therefore extracted a genomic region of 41 nucleotides centered on this position for further analyses.

Filtering to remove unspecific "background" clusters for AGO and TNRC6

Because it is still possible that a substantial number of the clusters we obtained contain degradation products of abundantly expressed mRNAs and because a number of proteins that associate with the RISC complex have a molecular weight that is similar to that of AGO proteins and may be responsible for some of the sequence reads/clusters that we obtained in the experiment with FLAG-tagged AGO we have collected PAR-CLIP data for a number of proteins and identified the AGO-specific clusters as follows. We built similar clusters for all the proteins that we investigated (PUM2, QKI, IGF2BP1-3, AGO1-4, TNRC6A-C), we compared the clusters, and when two clusters bound by two different proteins overlapped by more than 75% of their total length we considered that the two proteins shared a cluster. Finally, we discarded the following AGO clusters: clusters in which no position had a T to C mutation rate greater than 0.2, the experimentally determined T to C mutation rate at non-crosslinked sites; clusters that were shared between AGO libraries and libraries of other RBPs, with the number of sequence reads in the AGO libraries being less than 1/10 of the number of sequence reads in the other library. After applying these filters we obtained 17,319 AGO1-4 binding regions. We applied the same procedure to the clusters that we obtained from miR-124 and miR-7 transfection experiments.

Analysis of crosslinked position with respect to miRNA seed-complementary sequence

We identified all the target regions (T to C anchored regions of 41 nucleotides) that have an 8-mer (A opposite miRNA position 1 and perfect match at miRNA positions 2-8) seed match and we extended symmetrically the seed-complementary region by 20 nt to the left and right. We then computed the positional T to C mutation frequency in these regions and normalized it over the length of the target region.

Identification of pairing regions of miRNAs within CCRs

To determine whether positions other than the seed region may be involved in base-pairing interaction with targets, we first took the T to C anchored target regions and identified those that had at least a 6-mer (2-6 and A opposite miRNA position 1, 2-7 or 3-8) seed complementarity to at least one of the top 100 most expressed miRNAs in HEK293 cells. For each of these T to C anchored regions and each miRNA that matched to it, we identified all the occurrences of complementarities

of at least 4 nucleotides between the miRNA and the putative target region. Each of these was counted with a weight 1/n towards the positional profile of miRNA-target site matches, with n being the number of miRNAs that matched the putative target region.

Analysis of transcript stabilization as a function of the type of miRNA binding sites

We constructed the distribution of log-fold-changes of transcripts with various types of PAR-CLIP clusters, and we compared them with the distribution of log-fold-changes of transcripts that did not yield PAR-CLIP clusters, although they were expressed, as determined by the microarray measurements. The categories of transcripts were the following:

1. Transcripts with various types of miRNA seed matches
    - At most 6-mer match: 1-6 (with A opposite miRNA position 1), 2-7, 3-8, 4-9 match to at least one of the top 100 most abundant miRNAs.
    - At most 7-mer match: 1-7 (with A opposite miRNA position 1), 2-8, 3-9 match to at least one of the top 100 most abundant miRNAs.
    - At most 8-mer match: 1-8 (with A opposite miRNA position 1), 2-9 match to at least one of the top 100 most abundant miRNAs.
    - At most 9-mer match: 1-9 (with A opposite miRNA position 1) match to at least one of the top 100 most abundant miRNAs.
2. Transcripts with PAR-CLIP clusters originating exclusively in a particular transcript region (5'UTR, CDS, 3'UTR).
3. Transcripts with 1, 2, 3, 4 or more non-overlapping PAR-CLIP clusters.

Digital Gene Expression (DGE)

The sequence reads from the DGE (Illumina) experiments have been analyzed in a manner similar to that described above in the section "Adapter removal and sequence annotation". We only considered genomic and transcript matches containing the GATC recognition sequence of the DpnII restriction enzyme directly upstream of the mapped sequence read. For our analyses we further used sequence reads that had a perfect match in the genome. The probability that a sequence read originates in a given locus was then computed as 1/n of loci to which the sequence read can be mapped. The sequence reads were also mapped to the mRNA sequences and then we computed an expression level per gene. This was defined as the sum of the weighted copies of all sequence reads that can be mapped to transcripts that originate in that gene. Finally, to assess the accuracy of the expression level measurements, we correlated the logarithm of the expression level measured Affymetrix GeneChip microarray with the logarithm expression level measured using the DGE technology. The Spearman correlation coefficient was 0.68. We found a considerable number of transcripts

that could be detected by sequencing (22,465) and that were undetectable on the microarrays (on which we measured 16,063 transcripts). Correlation between biological replicates of HEK293 cells was higher than 0.99.

Analysis of miRNA-induced destabilization of crosslinked and non-crosslinked miR-124 and miR-7 targets

We intersected the transcripts with the background-noise-filtered PAR-CLIP clusters obtained after miR-124 and miR-7 transfection (see âFiltering to remove unspecific âbackgroundâ clusters for AGO and TNRC6â section above) with those for which we had destabilization and AGO-IP Affymetrix microarray measurements. We then constructed, for each miRNA, three non-overlapping sets of transcripts: those with PAR-CLIP clusters exclusively in the 3'UTR, with PAR-CLIP clusters exclusively in the CDS, and transcripts that did not yield any PAR-CLIP clusters. For each set, we computed the average log2 fold change upon miRNA transfection, and the average log2 fold enrichment in the AGO-IP. We compared these values between transcripts with and transcripts without PAR-CLIP clusters (Figure 59). The error bars on the bar plot represent 95% confidence intervals on the mean log2 fold changes. Finally, we performed Wilcoxon's rank sum test to assess the significance of the difference in the log2 fold changes of pairs of transcript sets. We also looked at various combinations of CLIP cluster locations (Figure 59) that occurred more than 25 times in a given data set. Finally, we compared the destabilization and AGO-binding of crosslinked and non-crosslinked single miR-124 and miR-7 seed matches (Figure 59). A seed match was defined as a match to nucleotides 1-7, 2-8 or 1-8 of the miRNA (both miRNAs start with U, so a 1-7 or 1-8 seed match also means having an A opposite nucleotide 1 of the miRNA). A seed match was considered "crosslinked" if it overlapped with a CLIP cluster from the corresponding transfection library.

Estimation of miRNA expression based on SOLEXA sequencing

The miRNA profile was generated from Solexa sequencing runs containing small RNAs from the following libraries: AGO1- IP and lysates of AGO1-4 IP, which were combined and denoted lysate in Figure 15C. The miRNA annotation was preformed as described in Berninger et al. [16], Landgraf et al. [129].

Plots of motif frequency versus enrichment

We performed a 7-mer word enrichment analysis based on the T to C anchored target regions from the miRNA transfection experiments. We enumerated all words of length 7 and we determined their frequency in the real set as well as in a background set of shuffled sequences with the same dinucleotide content. For each 7-mer, we then calculated its enrichment as the ratio of the two frequencies. Additionally, we calculated for each 7-mer the posterior probability that the frequency of the 7-mer is different in foreground and background allowing for sampling noise [16]. To determine whether the enriched motifs may correspond to miRNAs, all significantly enriched motifs (with a poste-

rior $\geqslant$ 0.99) were aligned with Needleman-Wunsch algorithm (penalties: gapopening -4, gapextension -4) to the reverse complement of the transfected and to the top 20 most expressed in HEK293 miRNAs. We only reported cases in which the enriched word mapped with 0 or 1 errors to the first 9 positions of one of these miRNAs.

Identification of significantly enriched types of miRNA binding sites

In order to identify individual miRNA binding sites in the sequence data we first defined a set of putative "binding models". These were either contiguous matches to at least 6 nucleotides of a miRNA, or matches that had a single structural defect. This was defined as either an internal loop or a bulge either in the miRNA or in the mRNA. For each of the 553 miRNAs we enumerated all these binding models, and we determined the enrichment of the T to C anchored regions in each of these models, relative to the average over 10 dinucleotide randomized sequence sets. Using a cutoff of 10^{-20} in the probability that the real set had a lower frequency of occurrence compared to the randomized sets, which we used as a measure of the significance of the enrichment, we found all the T to C anchored regions that contained at least one significantly enriched binding model from one of the top 100 most expressed miRNAs within 10 nucleotides of the T to C mutation site. To obtain a comprehensive list of target sites we added to these the 7-mer nucleotide matches (within the same 10 nucleotides of the T to C mutation) to positions 1-7 or 2-8 of one of the top 100 most expressed miRNAs, irrespective of whether the T to C anchored regions were enriched in these 7-mers.

Correlation of miRNA seed family expression with frequencies of occurrence of seed-complementary motif

From all samples of smirnadb [129], all miRNAs that had at least 50 counts in total from all samples were used to build seed groups (defined by the motif found at positions 2-8). We added an additional sample, which was generated by pooling together the miRNA reads from deep sequencing of HEK293 small RNA as well as AGO1-4 IPs without crosslinking. For each sample, we computed the expression of a seed group as the sum of the sequence reads of all miRNAs that were part of the seed group. We correlated the seed expression with the frequency of the seed-complementary motif in the T to C anchored regions.

Co-occurrence of miRNA seed pairs within CCRs

To determine if the crosslinked regions are enriched in pairs of binding sites for highly expressed miRNAs. Assuming that not all of these sites may have been captured in our experiment, we used for this purpose the 17,319 cluster regions that we extended by 32 nucleotides on either side. We scanned these regions for non-overlapping 7-mers corresponding to the positions 2-8 of the top 20 most expressed miRNAs in HEK293 cells. We performed a similar procedure using 100 randomized variants of the extended clusters that preserved the dinucleotide composition. As additional controls we performed, first, the same pro-

cedure using 20 randomly selected miRNAs (Figure 58F) and secondly counting of the number of seed match pair occurrence in the extended clusters for 100 sets of 20 randomly selected miRNAs (Figure 58H). A visualization of seed match pair occurrence is shown in Figure 58G.

Properties of crosslinked and non-crosslinked miRNA seed matches

For the analyses whose results are presented in Figure 59, we needed to intersect the CLIP transcript sets with the transcript set measured by the Affymetrix microrray. In order to study the properties of crosslinked and predicted but non-crosslinked seed complementary matches we do not need to make this intersection, and we therefore considered the entire set of miRNA seed matches that are present in the representative RefSeq transcripts. We chose as the representative RefSeq transcript for a given gene that transcript that had the median 3'UTR length from all RefSeq transcripts corresponding to a gene. RefSeq transcripts that could not be detected in the DGE transcriptome profiling were discarded. For the analysis of the miR-124 and miR-7 transfection libraries, we scanned the 5'UTR, CDS and 3'UTRs of representative expressed RefSeq transcripts for 7-mer or 8-mer seed matches to miR-124 or miR-7, and intersected these with the background-noise-filtered miR-124 and miR-7 PAR-CLIP clusters to identify the crosslinked and non-crosslinked seed matches. In parallel, we scanned the 5'UTR, CDS and 3'UTRs of representative expressed RefSeq transcripts for 7-mer and 8-mer seed matches to miR-15, miR-20, miR-103, miR-19, let-7 representing the top expressed miRNA families in HEK293 cells. These seed matches were then separated into crosslinked and non-crosslinked based on the intersection with the background-noise-filtered AGO1-4, PAR-CLIP clusters. Furthermore, because we wanted to analyze properties of the environment of the putative miRNA target sites, we only considered seed matches located at least 100 nucleotides away from either of the boundaries of the transcript. For each individual seed match, we computed the following quantities:

SELECTION PRESSURE is the posterior probability that a seed complementary region is under evolutionary selection pressure, as computed by the ElMMo algorithm described in Gaidatzis et al. [70].

PREDICTED DESTABILIZATION SCORE is a score that characterizes the extent to which the environment of a seed match is favorable for its functionality in mRNA destabilization, as computed by the TargetScanS method [83]. For the analysis, we downloaded the TargetScan 5.1 from the `www.TargetScan.org` website.

LOCAL AU CONTENT is the proportion of A + U nucleotides within 50 nucleotides upstream and 50 nucleotides downstream of the miRNA binding site, defined as a 20 nt-long region, anchored at the 3âend by the seed-matching region.

TARGET SITE EOPEN is similarly defined in terms of the energy required to open the secondary structure of the target in a region of 20 nucleotides anchored at the 3' end by the seed-complementary region (opposite positions 1-8 of the miRNA). This was computed using the program RNAup of the Vienna package [102]

with the following parameters: u=20 (length of the window required to be single-stranded), w=50 (maximal length of the interacting region). The rest of the parameters were left with their default values. The negative value of this energy can be viewed as a measure of accessibility.

We tested whether the four properties introduced above took significantly different values when comparing crosslinked to non-crosslinked seed matches using Wilcoxon's rank sum test.

Codon adaptation index around crosslinked and non-crosslinked seed matches

We compared the Codon Adaptation Index (CAI) [193] around crosslinked and non-crosslinked seed matches as follows. We estimated an optimal human codon usage by analyzing all the CDS from the 25% highest expressed genes among all the genes expressed in at least one of the two "whole brain" samples of the SymAtlas project [205]. This set of genes was determined by reanalyzing the two Affymetrix CEL files using the pipeline described above in the 'Analysis of miRNA knockdown and overexpression experiments' section. We then anchored all sequences at the codon covering the 5' end of seed match (1-7, 2-8, or 1-8 of miR-15, miR-20, miR-103, miR-19, let-7 miRNAs) and computed the CAI for the 70 codons upstream and downstream of the anchor, i.e. a total of 141 codons. The 7-mer or 8-mer seed match is entirely covered by codons 0, 1 and 2, which highly constrains the codon usage at these positions, making it uninformative. The figure therefore does not show the CAI at these positions. For crosslinked seed matches, we smoothed the profile using a moving average of 5.

Analysis of positional bias of crosslinked and non-crosslinked regions

We set to determine whether crosslinked seed matches (1-7, 2-8, or 1-8 of miR-15, miR-20, miR-103, miR-19, let-7 miRNAs) have a positional bias relative to the STOP codon. Noting that at least in the 4 AGO PAR-CLIP libraries, crosslinked seed matches tended to be located in CDS of shorter lengths than their non-crosslinked counterparts, we performed local polynomial regression [40], fitting the distance between the seed matches and the STOP codon to the CDS length (Figure 59M,N). The loess fit and 95% confidence interval on the distance to the STOP codon given the CDS length were obtained using R's loess and predict loess functions with default parameters. The miRNA transfection and AGO PAR-CLIP libraries were separately analyzed, and loess fits were computed separately for crosslinked and non-crosslinked seed matches (Figure 59K-N, shown in red and black, respectively). Finally, we represented the expected distance to the STOP codon as a function of the CDS length assuming that seed matches are distributed uniformly over the CDS (dashed blue curve). We used the same methodology to determine whether crosslinked sites are located preferentially towards a 3'UTR boundary (stop-codon or polyA-tail) instead of the stop-codon.

Comparison of the set of targets determined by the experimental assay (PAR-CLIP) and computational methods (ElMMo, TargetScan 5.1)

We computed the number of seed matches to each of the top 5 expressed miRNA families in the top 1000 CCRs from the AGO-PAR-CLIP. For each of these 5 miRNA families, we selected an equal number of target sites predicted by the ElMMo method, located on the mRNAs that could be detected in the DGE expression profiling (i.e. with at least one tag count), and starting from targets predicted with highest confidence. In addition, only genes that are expressed above the median on the arrays (i.e., the arrays in which the miRNAs are inhibited or not present) were considered in the analysis. We repeated the procedure using the TargetScan context scores, TargetScan PCT and Pictar. The ElMMo and TargetScan miRNA prediction methods only scan the mRNA 3'UTRs for target sites. Therefore, we determined a fourth set of miRNA target sites through keeping only the CCRs harboring a seed match to at least one of the top 5 miRNA families, and located in the 3'UTR region of an mRNA. Finally, for each of these 6 sets of miRNA targets and each of the top 5 miRNA families, we determined the average log2 fold change in gene expression upon transfecting the antisense 2'-O-methyl oligonucleotide cocktail as well as the 95% confidence interval on the mean log2 fold change. We performed the same analysis on the miR-7 and miR-124 transfection data sets, this time analyzing only CCRs containing seed matches to miR-7 or miR-124.

Stability of transcripts containing CCRs with 6-mer seed complementary matches

For all mRNAs representative of genes detected through DGE profiling, we computed the number of 3'UTR-located 6-mer and 7-mer (or longer) seed matches to the top 5 expressed miRNA families. We then plotted the mean log2 fold change in gene expression following the transfection of the antisense 2'-O-methyl oligonucleotide cocktail as a function of the number of 6-mer and 7-mer (or better) seed matches, as well as the 95% confidence interval on the mean log2 fold change. Finally, we performed the same analysis on the miR-7 and miR-124 transfection data sets, this time analyzing only seed matches to miR-7 and miR-124.

BIBLIOGRAPHY

[1] W Akamatsu, H Fujihara, T Mitsuhashi, M Yano, S Shibata, Y Hayakawa, H J Okano, S Sakakibara, H Takano, T Takano, T Takahashi, T Noda, and H Okano. The RNA-binding protein HuD regulates neuronal cell identity and maturation. *Proceedings of the National Academy of Sciences of the United States of America*, 102:4625–4630, 2005.

[2] B Alberts, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. *Molecular Biology of the Cell*. Garland Science, 4th edition, 2002. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Molecular+Biology+of+the+Cell#2.

[3] Victor Ambros. The functions of animal microRNAs. *Nature*, 431(7006):350–355, 2004. URL http://www.nature.com/nature/journal/v431/n7006/abs/nature02871.html.

[4] Stefan Ludwig Ameres, Javier Martinez, and Renée Schroeder. Molecular basis for target RNA recognition and cleavage by human RISC. *Cell*, 130(1):101–12, July 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.04.037. URL http://www.ncbi.nlm.nih.gov/pubmed/17632058.

[5] A. Andrus and R.G. Kuimelis. *Base composition analysis of nucleosides using HPLC*, chapter 10. 2001.

[6] Alexei A Aravin, Mariana Lagos-Quintana, Abdullah Yalcin, Mihaela Zavolan, Debora Marks, Ben Snyder, Terry Gaasterland, Jutta Meyer, and Thomas Tuschl. The small RNA profile during Drosophila melanogaster development. *Developmental cell*, 5(2): 337–50, August 2003. ISSN 1534-5807. URL http://www.ncbi.nlm.nih.gov/pubmed/12919683.

[7] Aaron Arvey, Erik Larsson, Chris Sander, Christina S Leslie, and Debora S Marks. Target mRNA abundance dilutes microRNA and siRNA activity. *Molecular systems biology*, 6:363, April 2010. ISSN 1744-4292. doi: 10.1038/msb.2010.24. URL http://dx.doi.org/10.1038/msb.2010.24.

[8] Daehyun Baek, J. Villén, Chanseok Shin, F.D. Camargo, S.P. Gygi, and David P. Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, 2008. doi: 10.1038/nature07242. URL http://www.nature.com/nature/journal/v455/n7209/abs/nature07242.html.

[9] Shveta Bagga, John Bracht, Shaun Hunter, Katlin Massirer, Janette Holtz, Rachel Eachus, and Amy E Pasquinelli. Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell*, 122(4):553–63, 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2005.07.031. URL http://www.ncbi.nlm.nih.gov/pubmed/16122423.

[10] Samuel Bandara, Johannes P Schlöder, Roland Eils, Hans Georg Bock, and Tobias Meyer. Optimal experimental design for parameter estimation of a cell signaling model. *PLoS computational biology*, 5(11):e1000558, November 2009. ISSN 1553-7358.

doi: 10.1371/journal.pcbi.1000558. URL http://dx.plos.org/10.1371/journal.pcbi.1000558.

[11] J Bard, S Rhee, and M Ashburner. An ontology for cell types. *Genome Biology*, 6:R21, 2005. URL http://www.biomedcentral.com/1465-6906/6/R21.

[12] David P. Bartel. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, 116(2):281–297, 2004. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867404000455.

[13] David P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136:215–233, 2009.

[14] Derek W Bartlett and Mark E Davis. Insights into the kinetics of siRNA-mediated gene silencing from live-cell and live-animal bioluminescent imaging. *Nucleic acids research*, 34(1):322–33, January 2006. ISSN 1362-4962. doi: 10.1093/nar/gkj439. URL http://nar.oxfordjournals.org/cgi/content/abstract/34/1/322.

[15] Simon Bennett. Solexa Ltd. *Pharmacogenomics*, 5(4):433–8, June 2004. ISSN 1462-2416. doi: 10.1517/14622416.5.4.433. URL http://www.ncbi.nlm.nih.gov/pubmed/15165179.

[16] P Berninger, D Gaidatzis, E Van Nimwegen, and M Zavolan. Computational analysis of small RNA cloning data. *Methods*, 44:13–21, 2008. URL http://linkinghub.elsevier.com/retrieve/pii/S1046202307001764.

[17] Suvendra N Bhattacharyya, Regula Habermacher, Ursula Martine, Ellen I Closs, and Witold Filipowicz. Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell*, 125(6):1111–24, June 2006. ISSN 0092-8674. doi: 10.1016/j.cell.2006.04.031. URL http://www.ncbi.nlm.nih.gov/pubmed/16777601.

[18] T Bieri, D Blasiar, P Ozersky, and I Antoshechkin. WormBase: new content and better access. *Nucleic acids research*, 35:D506–D510, 2006. URL http://nar.oxfordjournals.org/content/35/suppl_1/D506.full.

[19] T Bock, B Pakkenberg, and K Buschard. Increased islet volume but unchanged islet number in ob/ob mice. *Diabetes*, 52:1716–1722, 2003.

[20] S Bonner-Weir, D F Trent, and G C Weir. Partial pancreatectomy in the rat and subsequent defect in glucose-induced insulin release. *J Clin Invest*, 71:1544–1553, 1983.

[21] M I Borelli, M Rubio, M E Garcia, L E Flores, and J J Gagliardino. Tyrosine hydroxylase activity in the endocrine pancreas: changes induced by short-term dietary manipulation. *BMC Endocr Disord*, 3:2, 2003.

[22] B Boyerinas, S.-M. Park, N Shomron, M M Hedegaard, J Vinther, J S Andersen, C Feig, J Xu, C B Burge, and M E Peter. Identification of Let-7-Regulated Oncofetal Genes. *Cancer Res.*, 68:2587–2591, 2008.

[23] M Braun, R Ramracheya, M Bengtsson, Q Zhang, J Karanauskaite, C Partridge, P R Johnson, and P Rorsman. Voltage-gated ion channels in human pancreatic {beta}-cells Electrophysiological characterization and role in insulin secretion. *Diabetes*, 57:1618–1628, 2008.

[24] Julius Brennecke, Alexander Stark, Robert B Russell, and Stephen M Cohen. Principles of microRNA-target recognition. *PLoS biology*, 3(3):e85, March 2005. ISSN 1545-7885. doi: 10.1371/journal.pbio.0030085. URL http://dx.plos.org/10.1371/journal.pbio.0030085.

[25] R J Britten and E H Davidson. Gene regulation for higher cells: a theory. *Science*, 165(891):349–57, July 1969. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/5789433.

[26] Kevin Brown and James Sethna. Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E*, 68(2):1–9, 2003. ISSN 1063-651X. doi: 10.1103/PhysRevE.68.021904. URL http://link.aps.org/doi/10.1103/PhysRevE.68.021904.

[27] L Buckbinder, S Velasco-Miguel, Y Chen, N Xu, R Talbott, L Gelbert, J Gao, B R Seizinger, J S Gutkind, and N Kley. The p53 tumor suppressor targets a novel regulator of G protein signaling. *Proceedings of the National Academy of Sciences of the United States of America*, 94:7868–7872, 1997.

[28] R R Burgess, A A Travers, J J Dunn, and E K Bautz. Factor stimulating transcription by RNA polymerase. *Nature*, 221(5175):43–6, January 1969. ISSN 0028-0836. URL http://www.ncbi.nlm.nih.gov/pubmed/4882047.

[29] A.G. Busetto and J.M. Buhmann. Stable Bayesian Parameter Estimation for Biological Dynamical Systems. In *2009 International Conference on Computational Science and Engineering*, pages 148–157. IEEE, 2009. ISBN 978-1-4244-5334-4. doi: 10.1109/CSE.2009.134. URL http://www.computer.org/portal/web/csdl/doi/10.1109/CSE.2009.134.

[30] George Adrian Calin, Calin Dan Dumitru, Masayoshi Shimizu, Roberta Bichi, Simona Zupo, Evan Noch, Hansjuerg Aldler, Sashi Rattan, Michael Keating, Kanti Rai, Laura Rassenti, Thomas Kipps, Massimo Negrini, Florencia Bullrich, and Carlo M Croce. Frequent deletions and down-regulation of microRNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24):15524–9, November 2002. ISSN 0027-8424. doi: 10.1073/pnas.242606799. URL http://www.pnas.org/cgi/content/abstract/99/24/15524.

[31] George Adrian Calin, Chang-Gong Liu, Cinzia Sevignani, Manuela Ferracin, Nadia Felli, Calin Dan Dumitru, Masayoshi Shimizu, Amelia Cimmino, Simona Zupo, Mariella Dono, Marie L Dell'Aquila, Hansjuerg Alder, Laura Rassenti, Thomas J Kipps, Florencia Bullrich, Massimo Negrini, and Carlo M Croce. MicroRNA profiling reveals distinct signatures in B cell chronic

lymphocytic leukemias. *Proceedings of the National Academy of Sciences of the United States of America*, 101(32):11755–60, August 2004. ISSN 0027-8424. doi: 10.1073/pnas.0404432101. URL http://www.ncbi.nlm.nih.gov/pubmed/15284443.

[32] Cristian I. Castillo-Davis and Daniel L. Hartl. GeneMerge–post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19(7):891–892, May 2003. ISSN 14602059. doi: 10.1093/bioinformatics/btg114. URL http://www.bioinformatics.oupjournals.org/cgi/doi/10.1093/bioinformatics/btg114.

[33] Ho-Man Chan, Lai-Sheung Chan, Ricky Ngok-Shun Wong, and Hung-Wing Li. Direct Quantification of Single-Molecules of MicroRNA by Total Internal Reflection Fluorescence Microscopy. *Analytical Chemistry*, 82:6911–6918, July 2010. ISSN 0003-2700. doi: 10.1021/ac101133x. URL http://dx.doi.org/10.1021/ac101133x.

[34] Chris Cheadle, Jinshui Fan, Yoon S Cho-Chung, Thomas Werner, Jill Ray, Lana Do, Myriam Gorospe, and Kevin G Becker. Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability. *BMC genomics*, 6(1):75, January 2005. ISSN 1471-2164. doi: 10.1186/1471-2164-6-75. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1156890&tool=pmcentrez&rendertype=abstract.

[35] Caifu Chen, Dana a Ridzon, Adam J Broomer, Zhaohui Zhou, Danny H Lee, Julie T Nguyen, Maura Barbisin, Nan Lan Xu, Vikram R Mahuvakar, Mark R Andersen, Kai Qin Lao, Kenneth J Livak, and Karl J Guegler. Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic acids research*, 33(20):e179, January 2005. ISSN 1362-4962. doi: 10.1093/nar/gni178. URL http://www.ncbi.nlm.nih.gov/pubmed/16314309.

[36] C A Chenard and S Richard. New implications for the QUAKING RNA binding protein in human disease. J. Neurosci. *J Neurosci Res*, 86:233–242, 2008.

[37] E C Cheung, N Joza, N A Steenaart, K A McClellan, M Neuspiel, S McNamara, J G MacLaurin, P Rippstein, D S Park, G C Shore, H M McBride, J M Penninger, and R S Slack. Dissociating the dual roles of apoptosis-inducing factor in maintaining mitochondrial structure and apoptosis. *Embo J*, 25:4061–4073, 2006.

[38] Sung Wook Chi, Julie B Zang, Aldo Mele, and Robert B Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–86, July 2009. ISSN 1476-4687. doi: 10.1038/nature08170. URL http://dx.doi.org/10.1038/nature08170.

[39] AG Clark, MB Eisen, DR Smith, CM Bergman, and B. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450:203—218, 2007. URL http://www.nature.com/nature/journal/v450/n7167/abs/nature06341.html.

[40] W.S. Cleveland, E. Grosse, and W.M. Shyu. *Local regression models*. Wadsworth & Brooks / Cole, 1992.

[41] A W Cohen, D S Park, S E Woodman, T M Williams, M Chandra, J Shirani, A de Souza, R N Kitsis, R G Russell, L M Weiss, B Tang, L A Jelicks, S M Factor, V Shtutin, H B Tanowitz, and M P Lisanti. Caveolin-1 null mice develop cardiac hypertrophy with hyperactivation of p42/44 MAP kinase in cardiac fibroblasts. *Am J Physiol Cell Physiol*, 284:C457–474, 2003.

[42] The Bovine Genome Sequencing Consortium, Analysis, Christine G. Elsik, Ross L. Tellam, and Kim C. Worley. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, 324:522—-528, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19390049.

[43] I de Silanes, M Zhan, A Lal, X Yang, and M Gorospe. Identification of a target RNA motif for RNA-binding protein HuR. *Proceedings of the National Academy of Sciences of the United States of America*, 101:2987–2992, 2004.

[44] A.P. Dempster, N.M. Laird, D.B. Rubin, and Others. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. URL http://www.ams.org/leavingmsn?url=http://links.jstor.org/sici?sici=0035-9246(1977)39:1<1:MLFIDV>2.0.CO;2-Z&origin=MSN.

[45] Euthymios Dimitriadis, Theoni Trangas, Stavros Milatos, Periklis G Foukas, Ioannis Gioulbasanis, Nelly Courtis, Finn C Nielsen, Nikos Pandis, Urania Dafni, Georgia Bardi, and Panayotis Ioannidis. Expression of oncofetal RNA-binding protein CRD-BP/IMP1 predicts clinical outcome in colon cancer. *International journal of cancer. Journal international du cancer*, 121(3):486–94, August 2007. ISSN 0020-7136. doi: 10.1002/ijc.22716. URL http://www.ncbi.nlm.nih.gov/pubmed/17415713.

[46] Xavier C Ding and Helge Grosshans. Repression of C. elegans microRNA targets at the initiation level of translation requires GW182 proteins. *The EMBO journal*, 28(3):213–22, February 2009. ISSN 1460-2075. doi: 10.1038/emboj.2008.275. URL http://dx.doi.org/10.1038/emboj.2008.275.

[47] John G Doench and Phillip A Sharp. Specificity of microRNA target selection in translational repression. *Genes & development*, 18(5):504–11, March 2004. ISSN 0890-9369. doi: 10.1101/gad.1184404. URL http://www.ncbi.nlm.nih.gov/pubmed/15014042.

[48] Mary K Doherty, Dean E Hammond, Michael J Clague, Simon J Gaskell, and Robert J Beynon. Turnover of the human proteome: determination of protein intracellular stability by dynamic SILAC. *Journal of proteome research*, 8(1):104–12, January 2009. ISSN 1535-3893. doi: 10.1021/pr800641v. URL http://dx.doi.org/10.1021/pr800641v.

[49] G Dreyfuss, Y D Choi, and S A Adam. Characterization of heterogeneous nuclear RNA-protein complexes in vivo with monoclonal antibodies. *Molecular and Cellular Biology.*, 4:1104–1114, 1984.

[50] Anja M Duursma, Martijn Kedde, Mariette Schrier, Carlos le Sage, and Reuven Agami. miR-148 targets human DNMT3b protein coding region., 2008. ISSN 1469-9001. URL http://www.ncbi.nlm.nih.gov/pubmed/18367714.

[51] W S Dynan and R Tjian. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell*, 35(1):79–87, November 1983. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/6313230.

[52] George Easow, Aurelio A Teleman, and Stephen M Cohen. Isolation of microRNA targets by miRNP immunopurification. *RNA*, 13(8):1198–204, August 2007. ISSN 1355-8382. doi: 10.1261/rna.563707. URL http://www.ncbi.nlm.nih.gov/pubmed/17592038.

[53] M.S. Ebert, J.R. Neilson, and P.A. Sharp. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nature methods*, 4(9):721–726, 2007. doi: 10.1038/NMETH1079. URL http://www.nature.com/nmeth/journal/v4/n9/abs/nmeth1079.html.

[54] H. Edlund. Pancreatic organogenesis–developmental mechanisms and implications for therapy. *Nature Reviews Genetics*, 3: 524–532, 2002.

[55] RM Eglen. Muscarinic receptor subtypes in neuronal and nonâneuronal cholinergic function. *Autonomic and Autacoid Pharmacology*, 26:219—-233, 2006. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1474-8673.2006.00368.x/full.

[56] L Eliasson, X Ma, E Renstrom, S Barg, P O Berggren, J Galvanovskis, J Gromada, X Jing, I Lundquist, A Salehi, S Sewing, and P Rorsman. SUR1 regulates PKA-independent cAMP-induced granule priming in mouse pancreatic B-cells. *J Gen Physiol*, 121:181–197, 2003.

[57] Ran Elkon and Reuven Agami. Removal of AU bias from microarray mRNA expression data enhances computational identification of active microRNAs. *PLoS computational biology*, 4(10):e1000189, January 2008. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000189. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2533120&tool=pmcentrez&rendertype=abstract.

[58] E Englesberg, J Irr, J Power, and N Lee. Positive control of enzyme synthesis by gene C in the L-arabinose system. *Journal of bacteriology*, 90(4):946–57, October 1965. ISSN 0021-9193. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=315760&tool=pmcentrez&rendertype=abstract.

[59] Anton J Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S Marks. MicroRNA targets in Drosophila. *Genome biology*, 5(1):R1, January 2003. ISSN 1465-6914. doi: 10.1186/gb-2003-5-1-r1. URL http://www.ncbi.nlm.nih.gov/pubmed/14709173.

[60] A. Eulalio, J. Rehwinkel, M. Stricker, E. Huntzinger, S.F. Yang, T. Doerks, S. Dorner, Peer Bork, M. Boutros, and E. Izaurralde. Target-specific requirements for enhancers of decapping

in miRNA-mediated gene silencing. *Genes & development*, 21 (20):2558, 2007. URL http://genesdev.cshlp.org/cgi/content/abstract/21/20/2558.

[61] Ana Eulalio, Isabelle Behm-Ansmant, Daniel Schweizer, and Elisa Izaurralde. P-body formation is a consequence, not the cause, of RNA-mediated gene silencing. *Molecular and cellular biology*, 27(11):3970–81, 2007. ISSN 0270-7306. doi: 10.1128/MCB.00128-07. URL http://www.ncbi.nlm.nih.gov/pubmed/17403906.

[62] Ana Eulalio, Eric Huntzinger, and Elisa Izaurralde. Getting to the root of miRNA-mediated gene silencing. *Cell*, 132(1):9–14, January 2008. ISSN 0092-8674. doi: 10.1016/j.cell.2007.12.024. URL http://www.ncbi.nlm.nih.gov/pubmed/18191211.

[63] Ana Eulalio, Eric Huntzinger, Tadashi Nishihara, Jan Rehwinkel, Maria Fauser, and Elisa Izaurralde. Deadenylation is a widespread effect of miRNA regulation. *RNA*, 15(1):21–32, 2009. ISSN 1469-9001. doi: 10.1261/rna.1399509. URL http://www.ncbi.nlm.nih.gov/pubmed/19029310.

[64] M.R. Fabian, G. Mathonnet, T. Sundermeier, H. Mathys, J.T. Zipprich, Y.V. Svitkin, F. Rivas, M. Jinek, J. Wohlschlegel, J.A. Doudna, and Others. Mammalian miRNA RISC recruits CAF1 and PABP to affect PABP-dependent deadenylation. *Molecular cell*, 35:868–880, 2009. URL http://linkinghub.elsevier.com/retrieve/pii/S1097276509005504.

[65] A Favre, G Moreno, M O Blondel, J Kliber, F Vinzens, and C Salet. 4-thiouridine photosensitized RNA-protein crosslinking in mammalian cells. Biochem. Biophys. Res. *Commun.*, 141: 847–854, 1986.

[66] Witold Filipowicz, S.N. Bhattacharyya, and N. Sonenberg. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Reviews Genetics*, 9(2):102–114, 2008. URL http://www.nature.com/nrg/journal/vaop/ncurrent/full/nrg2290.html.

[67] Joshua J Forman, Aster Legesse-Miller, and Hilary A Coller. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 105(39):14879–84, September 2008. ISSN 1091-6490. doi: 10.1073/pnas.0803230105. URL http://www.ncbi.nlm.nih.gov/pubmed/18812516.

[68] Caroline C Friedel, Lars Dölken, Zsolt Ruzsics, Ulrich H Koszinowski, and Ralf Zimmer. Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic acids research*, 37(17):e115, September 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp542. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2761256&tool=pmcentrez&rendertype=abstract.

[69] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P. Bartel. Most mammalian mRNAs are conserved

targets of microRNAs. *Genome research*, 19(1):92–105, 2009. ISSN 1088-9051. doi: 10.1101/gr.082701.108. URL http://genome.cshlp.org/cgi/content/abstract/19/1/92.

[70] Dimos Gaidatzis, Erik van Nimwegen, Jean Hausser, and Mihaela Zavolan. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC bioinformatics*, 8(1): 69, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-69. URL http://www.biomedcentral.com/1471-2105/8/69.

[71] A Galarneau and S Richard. Target RNA motif and target mRNAs of the Quaking STAR protein. Nat. Struct. Mol. *Biol.*, 12: 691–698, 2005.

[72] F Galbiati, D Volonte, J Liu, F Capozza, P G Frank, L Zhu, R G Pestell, and M P Lisanti. Caveolin-1 expression negatively regulates cell cycle progression by inducing G(0)/G(1) arrest via a p53/p21(WAF1/Cip1)-dependent mechanism. *Mol Biol Cell*, 12: 2229–2244, 2001.

[73] A Galgano, M Forrer, L Jaskiewicz, A Kanitz, M Zavolan, and A P Gerber. Comparative Analysis of mRNA Targets for Human PUF-Family Proteins Suggests Extensive Interaction with the miRNA Regulatory System. *PLoS ONE*, 3:e3164, 2008.

[74] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, January 2004. ISSN 1465-6914. doi: 10.1186/gb-2004-5-10-r80. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=545600&tool=pmcentrez&rendertype=abstract.

[75] B. Gentner, G. Schira, A. Giustacchini, Mario Amendola, B.D. Brown, M. Ponzoni, and L. Naldini. Stable knockdown of microRNA in vivo by lentiviral vectors. *Nature methods*, 6(1):63–66, 2009. URL http://www.nature.com/nmeth/journal/vaop/ncurrent/full/nmeth.1277.html.

[76] A P Gerber, S Luschnig, M A Krasnow, P O Brown, and D Herschlag. Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster. *Proc. Nat. Acad Sci.*, 103:4487–4492, 2006.

[77] RA Gibbs, GM Weinstock, ML Metzker, DM Muzny, and EJ. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428:492—-521, 2004. URL http://www.nature.com/nature/journal/vaop/ncurrent/abs/nature02426.html.

[78] Antonio J Giraldez, Yuichiro Mishima, Jason Rihel, Russell J Grocock, Stijn Van Dongen, Kunio Inoue, Anton J Enright, and

Alexander F Schier. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, 312(5770):75–9, 2006. ISSN 1095-9203. doi: 10.1126/science.1122689. URL http://www.ncbi.nlm.nih.gov/pubmed/16484454.

[79] Eva Gottwein, Neelanjan Mukherjee, Christoph Sachse, Corina Frenzel, William H Majoros, Jen-Tsan A Chi, Ravi Braich, Muthiah Manoharan, Jürgen Soutschek, Uwe Ohler, and Bryan R Cullen. A viral microRNA functions as an orthologue of cellular miR-155. *Nature*, 450(7172):1096–9, December 2007. ISSN 1476-4687. doi: 10.1038/nature05992. URL http://www.ncbi.nlm.nih.gov/pubmed/18075594.

[80] Sander Granneman, Grzegorz Kudla, Elisabeth Petfalski, and David Tollervey. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 106(24):9613–8, June 2009. ISSN 1091-6490. doi: 10.1073/pnas.0901997106. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2688437&tool=pmcentrez&rendertype=abstract.

[81] J R Greenberg. Ultraviolet light-induced crosslinking of mRNA to proteins. Nucl. *Nucleic acids research*, 6:715–732, 1979.

[82] S Griffiths-Jones, RJ Grocock, Stijn Van Dongen, A Bateman, and Anton J Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research*, 34:D140–144, 2006. URL http://nar.oxfordjournals.org/content/34/suppl_1/D140.full.

[83] A Grimson, K.K.H. Farh, W.K. Johnston, P. Garrett-Engele, Lee P. Lim, and David P. Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*, 27(1):91–105, 2007. URL http://linkinghub.elsevier.com/retrieve/pii/S1097276507004078.

[84] A Grishok, A E Pasquinelli, D Conte, N Li, S Parrish, I Ha, D L Baillie, A Fire, G Ruvkun, and C C Mello. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, 106(1):23–34, July 2001. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/11461699.

[85] J Gromada, I Franklin, and CB Wollheim. alpha-Cells of the endocrine pancreas: 35 years of research but the enigma remains. *Endocr Rev*, 28:84–116, 2007.

[86] Shuo Gu, Lan Jin, Feijie Zhang, Peter Sarnow, and Mark A Kay. Biological basis for restriction of microRNA targets to the 3' untranslated region in mammalian mRNAs. *Nature structural & molecular biology*, 16(2):144–50, February 2009. ISSN 1545-9985. doi: 10.1038/nsmb.1552. URL http://www.ncbi.nlm.nih.gov/pubmed/19182800.

[87] S Guil and J F Caceres. The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nature structural & molecular biology*, 14:591, 2007.

[88] Huili Guo, Nicholas T. Ingolia, Jonathan S. Weissman, and David P. Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840, August 2010. ISSN 0028-0836. doi: 10.1038/nature09267. URL http://dx.doi.org/10.1038/nature09267.

[89] Markus Hafner, Pablo Landgraf, Janos Ludwig, Amanda Rice, Tolulope Ojo, Carolina Lin, Daniel Holoch, Cindy Lim, and Thomas Tuschl. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, 44(1): 3–12, January 2008. ISSN 1046-2023. doi: 10.1016/j.ymeth.2007.09.009. URL http://www.ncbi.nlm.nih.gov/pubmed/18158127.

[90] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010. ISSN 0092-8674. doi: 10.1016/j.cell.2010.03.009. URL http://dx.doi.org/10.1016/j.cell.2010.03.009.

[91] Molly Hammell, Dang Long, Liang Zhang, Andrew Lee, C Steven Carmack, Min Han, Ye Ding, and Victor Ambros. mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein â enriched transcripts. *Nature Methods*, 5(9), 2008. doi: 10.1038/NMETH.1247.

[92] Todd W Harris, Igor Antoshechkin, Tamberlyn Bieri, Darin Blasiar, Juancarlos Chan, Wen J Chen, Norie De La Cruz, Paul Davis, Margaret Duesbury, Ruihua Fang, Jolene Fernandes, Michael Han, Ranjana Kishore, Raymond Lee, Hans-Michael Müller, Cecilia Nakamura, Philip Ozersky, Andrei Petcherski, Arun Rangarajan, Anthony Rogers, Gary Schindelman, Erich M Schwarz, Mary Ann Tuli, Kimberly Van Auken, Daniel Wang, Xiaodong Wang, Gary Williams, Karen Yook, Richard Durbin, Lincoln D Stein, John Spieth, and Paul W Sternberg. WormBase: a comprehensive resource for nematode research. *Nucleic acids research*, 38(Database issue):D463–7, January 2010. ISSN 1362-4962. doi: 10.1093/nar/gkp952. URL http://nar.oxfordjournals.org/cgi/content/abstract/38/suppl_1/D463.

[93] RK Hartmann, A Bindereif, A Schön, and E Westhof, editors. *Handbook of RNA biochemistry*. Wiley-VCH, 1st edition, 2005. URL http://onlinelibrary.wiley.com/doi/10.1002/3527608192.fmatter/pdf.

[94] Jean Hausser, Philipp Berninger, Christoph Rodak, Yvonne Jantscher, Stefan Wirth, and Mihaela Zavolan. MirZ: an integrated microRNA expression atlas and target prediction resource. *Nucleic acids research*, 37(Web Server issue):W266–72, July 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp412. URL http://www.ncbi.nlm.nih.gov/pubmed/19468042.

[95] Jean Hausser, Markus Landthaler, Lukasz Jaskiewicz, Dimos Gaidatzis, and Mihaela Zavolan. Relative contribution of sequence and structure features to the mRNA binding of

Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome research*, 19(11):2009–20, November 2009. ISSN 1549-5469. doi: 10.1101/gr.091181.109. URL http://www.ncbi.nlm.nih.gov/pubmed/19767416.

[96] P Havlak, R Chen, KJ Durbin, A Egan, R Yanru, X.-Z. Song, G. M. Weinstock, and R. A. Gibbs. The Atlas genome assembly system. *Genome research*, 14:721–732, 2004. URL http://genome.cshlp.org/content/14/4/721.full.

[97] Lin He, J Michael Thomson, Michael T Hemann, Eva Hernando-Monge, David Mu, Summer Goodson, Scott Powers, Carlos Cordon-Cardo, Scott W Lowe, Gregory J Hannon, and Scott M Hammond. A microRNA polycistron as a potential human oncogene. *Nature*, 435(7043):828–33, 2005. ISSN 1476-4687. doi: 10.1038/nature03552. URL http://www.ncbi.nlm.nih.gov/pubmed/15944707.

[98] David G Hendrickson, Daniel J Hogan, Daniel Herschlag, James E Ferrell, and Patrick O. Brown. Systematic identification of mRNAs recruited to argonaute 2 by specific microRNAs and corresponding changes in transcript abundance. *PloS one*, 3(5):e2126, January 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0002126. URL http://www.ncbi.nlm.nih.gov/pubmed/18461144.

[99] David G Hendrickson, Daniel J Hogan, Heather L McCullough, Jason W Myers, Daniel Herschlag, James E Ferrell, and Patrick O. Brown. Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS biology*, 7(11):e1000238, November 2009. ISSN 1545-7885. doi: 10.1371/journal.pbio.1000238. URL http://dx.plos.org/10.1371/journal.pbio.1000238.

[100] S M Heywood and D S Kennedy. Purification of myosin translational control RNA and its interaction with myosin messenger RNA. *Biochemistry*, 15(15):3314–9, July 1976. ISSN 0006-2960. URL http://www.ncbi.nlm.nih.gov/pubmed/986157.

[101] LDW Hillier, W Miller, E Birney, W Warren, and RC. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432:695—-716, 2004. URL http://www.nature.com/nature/journal/v432/n7018/abs/nature03154.html.

[102] IL Hofacker. Vienna RNA secondary structure server. *Nucleic acids research*, 31:3429–3431, 2003. URL http://nar.oxfordjournals.org/content/31/13/3429.full.

[103] György Hutvágner and Phillip D Zamore. A microRNA in a multiple-turnover RNAi enzyme complex. *Science*, 297(5589):2056–60, October 2002. ISSN 1095-9203. doi: 10.1126/science.1073827. URL http://www.ncbi.nlm.nih.gov/pubmed/12154197.

[104] Fumiko Iwamoto, Michael Stadler, Katerina Chalupníková, Edward Oakeley, and Yoshikuni Nagamine. Transcription-dependent nucleolar cap localization and possible nuclear function of DExH RNA helicase RHAU. *Experimental cell research*,

314(6):1378–91, April 2008. ISSN 0014-4827. doi: 10.1016/j.yexcr.2008.01.006. URL http://www.ncbi.nlm.nih.gov/pubmed/18279852.

[105] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356, 1961.

[106] Qing Jing, Shuang Huang, Sabine Guth, Tyler Zarubin, Andrea Motoyama, Jianming Chen, Franco Di Padova, Sheng-Cai Lin, Hermann Gram, and Jiahuai Han. Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell*, 120(5):623–34, March 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2004.12.038. URL http://www.ncbi.nlm.nih.gov/pubmed/15766526.

[107] M K Joe, H J Lee, Y H Suh, K L Han, J H Lim, J Song, J K Seong, and M H Jung. Crucial roles of neuronatin in insulin secretion and high glucose-induced apoptosis in pancreatic beta-cells. *Cell Signal*, 20:907–915, 2008.

[108] J.D. Johnson, N.T. Ahmed, D.S. Luciani, Z.Q. Han, H. Tran, S. FujiMisler, H. Edlund, and K.S. Polonsky. Increased islet apoptosis in Pdx1(+/-) mice. *Journal of Clinical Investigation*, 111:1147–1160, 2003.

[109] A Kamal and K Datta. Upregulation of hyaluronan binding protein 1 (HABP1/p32/gC1qR) is associated with Cisplatin induced apoptosis. *Apoptosis*, 11:861–874, 2006.

[110] Fedor V Karginov, Cecilia Conaco, Zhenyu Xuan, Bryan H Schmidt, Joel S Parker, Gail Mandel, and Gregory J Hannon. A biochemical approach to identifying microRNA targets. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49):19291–6, December 2007. ISSN 1091-6490. doi: 10.1073/pnas.0709971104. URL http://www.ncbi.nlm.nih.gov/pubmed/18042700.

[111] D Karolchik, R M Kuhn, R Baertsch, G P Barber, H Clawson, M Diekhans, B Giardine, R A Harte, A S Hinrichs, F Hsu, K M Kober, W Miller, J S Pedersen, A Pohl, B J Raney, B Rhead, K R Rosenbloom, K E Smith, M Stanke, A Thakkapallayil, H Trumbower, T Wang, A S Zweig, D Haussler, and W J Kent. The UCSC Genome Browser Database: 2008 update. *Nucleic acids research*, 36(Database issue):D773–9, January 2008. ISSN 1362-4962. doi: 10.1093/nar/gkm966. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2238835&tool=pmcentrez&rendertype=abstract.

[112] M Kasahara, K Naruse, S Sasaki, Y Nakatani, W Qu, and B. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 446:714—-719, 2007. URL http://www.nature.com/nature/journal/v447/n7145/abs/nature05846.html.

[113] Martijn Kedde and Reuven Agami. Interplay between microRNAs and RNA-binding proteins determines developmental processes. *Cell Cycle*, 7(7):899–903, 2008.

[114] Martijn Kedde, Markus J Strasser, Bijan Boldajipour, Joachim A F Oude Vrielink, Krasimir Slanchev, Carlos le Sage, Remco

Nagel, P Mathijs Voorhoeve, Josyanne van Duijse, Ulf Andersson Ø rom, Anders H Lund, Anastassis Perrakis, Erez Raz, and Reuven Agami. RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. *Cell*, 131(7):1273–86, December 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.11.034. URL http://www.ncbi.nlm.nih.gov/pubmed/18155131.

[115] B L Kee. Id3 induces growth arrest and caspase-2-dependent apoptosis in B lymphocyte progenitors. *J Immunol*, 175:4518–4527, 2005.

[116] J D Keene. RNA regulons: coordination of post-transcriptional events. *Nature Reviews Genetics*, 8:533–543, 2007.

[117] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microRNA target recognition. *Nature genetics*, 39(10):1278–84, October 2007. ISSN 1546-1718. doi: 10.1038/ng2135. URL http://dx.doi.org/10.1038/ng2135.

[118] Aly A Khan, Doron Betel, Martin L Miller, Chris Sander, Christina S Leslie, and Debora S Marks. Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. *Nature biotechnology*, 27(6):549–55, June 2009. ISSN 1546-1696. doi: 10.1038/nbt.1543. URL http://dx.doi.org/10.1038/nbt.1543.

[119] Raya Khanin and Desmond J Higham. A Multi-step model for microRNA-mediated gene regulation. 2008.

[120] Raya Khanin and Veronica Vinciotti. Computational modeling of post-transcriptional gene regulation by microRNAs. *Journal of computational biology*, 15(3):305–16, April 2008. ISSN 1066-5277. doi: 10.1089/cmb.2007.0184. URL http://www.ncbi.nlm.nih.gov/pubmed/18333757.

[121] Y Kirino and Z Mourelatos. Site-specific crosslinking of human microRNPs to RNA targets. *RNA*, 14:2254–2259, 2008.

[122] W.P. Kloosterman, A.K. Lagendijk, R.F. Ketting, J.D. Moulton, and R.H. Plasterk. Targeted inhibition of miRNA maturation with morpholinos reveals a role for miR-375 in pancreatic islet development. *PLoS Biology*, 5:e203, 2007.

[123] Jan Krützfeldt, Nikolaus Rajewsky, Ravi Braich, Kallanthottathil G Rajeev, Thomas Tuschl, Muthiah Manoharan, and Markus Stoffel. Silencing of microRNAs in vivo with 'antagomirs'. *Nature*, 438(7068):685–9, December 2005. ISSN 1476-4687. doi: 10.1038/nature04303. URL http://dx.doi.org/10.1038/nature04303.

[124] R N Kulkarni, M Holzenberger, D Q Shih, U Ozcan, M Stoffel, M A Magnuson, and C R Kahn. beta-cell-specific deletion of the Igf1 receptor leads to hyperinsulinemia and glucose intolerance but does not alter beta-cell mass. *Nat Genet*, 31:111–115, 2002.

[125] M Kuramochi, H Fukuhara, T Nobukuni, T Kanbe, T Maruyama, H P Ghosh, M Pletcher, M Isomura, M Onizuka, T Kitamura,

T Sekiya, R H Reeves, and Y Murakami. TSLC1 is a tumor-suppressor gene in human non-small-cell lung cancer. *Nat Genet*, 27:427–430, 2001.

[126] Mariana Lagos-Quintana, R Rauhut, W Lendeckel, and T Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–8, 2001. ISSN 0036-8075. doi: 10.1126/science.1064921. URL http://www.ncbi.nlm.nih.gov/pubmed/11679670.

[127] Eric C Lai. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nature genetics*, 30(4):363–4, 2002. ISSN 1061-4036. doi: 10.1038/ng865. URL http://www.ncbi.nlm.nih.gov/pubmed/11896390.

[128] Sabbi Lall, Dominic Grün, Azra Krek, Kevin Chen, Yi-Lu Wang, Colin N Dewey, Pranidhi Sood, Teresa Colombo, Nicolas Bray, Philip Macmenamin, Huey-Ling Kao, Kristin C Gunsalus, Lior Pachter, Fabio Piano, and Nikolaus Rajewsky. A genome-wide map of conserved microRNA targets in C. elegans. *Current biology*, 16(5):460–71, 2006. ISSN 0960-9822. doi: 10.1016/j.cub.2006.01.050. URL http://dx.doi.org/10.1016/j.cub.2006.01.050.

[129] Pablo Landgraf, Mirabela Rusu, Robert Sheridan, Alain Sewer, Nicola Iovino, Alexei Aravin, Sébastien Pfeffer, Amanda Rice, Alice O Kamphorst, Markus Landthaler, Carolina Lin, Nicholas D Socci, Leandro Hermida, Valerio Fulci, Sabina Chiaretti, Robin Foà, Julia Schliwka, Uta Fuchs, Astrid Novosel, Roman-Ulrich Müller, Bernhard Schermer, Ute Bissels, Jason Inman, Quang Phan, Minchen Chien, David B Weir, Ruchi Choksi, Gabriella De Vita, Daniela Frezzetti, Hans-Ingo Trompeter, Veit Hornung, Grace Teng, Gunther Hartmann, Miklos Palkovits, Roberto Di Lauro, Peter Wernet, Giuseppe Macino, Charles E Rogler, James W Nagle, Jingyue Ju, F Nina Papavasiliou, Thomas Benzing, Peter Lichter, Wayne Tam, Michael J Brownstein, Andreas Bosio, Arndt Borkhardt, James J Russo, Chris Sander, Mihaela Zavolan, and Thomas Tuschl. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129(7):1401–14, 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.04.040. URL http://www.ncbi.nlm.nih.gov/pubmed/17604727.

[130] Markus Landthaler, Dimos Gaidatzis, Andrea Rothballer, Po Yu Chen, Steven Joseph Soll, Lana Dinic, Tolulope Ojo, Markus Hafner, Mihaela Zavolan, and Thomas Tuschl. Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. *RNA*, 14(12):2580–96, December 2008. ISSN 1469-9001. doi: 10.1261/rna.1351608. URL http://www.ncbi.nlm.nih.gov/pubmed/18978028.

[131] Rosalind C. Lee, Rhonda L. Feinbaum, and Victor Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75:843–854, 1993.

[132] Yoontae Lee, Chiyoung Ahn, Jinju Han, Hyounjeong Choi, Jaekwang Kim, Jeongbin Yim, Junho Lee, Patrick Provost, Olof Rå dmark, Sunyoung Kim, and V Narry Kim. The nuclear RNase

III Drosha initiates microRNA processing. *Nature*, 425(6956):415–9, October 2003. ISSN 1476-4687. doi: 10.1038/nature01957. URL http://www.ncbi.nlm.nih.gov/pubmed/14508493.

[133] Erel Levine, Zhongge Zhang, Thomas Kuhlman, and Terence Hwa. Quantitative characteristics of gene regulation by small RNA. *PLoS biology*, 5(9):e229, September 2007. ISSN 1545-7885. doi: 10.1371/journal.pbio.0050229. URL http://dx.plos.org/10.1371/journal.pbio.0050229.

[134] Benjamin P. Lewis, I-hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798, 2003. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867403010183.

[135] Benjamin P Lewis, Christopher B Burge, and David P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120 (1):15–20, 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2004.12.035. URL http://www.ncbi.nlm.nih.gov/pubmed/15652477.

[136] Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, Anthony C Schweitzer, John E Blume, Xuning Wang, Jennifer C Darnell, and Robert B Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–9, November 2008. ISSN 1476-4687. doi: 10.1038/nature07488. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2597294&tool=pmcentrez&rendertype=abstract.

[137] Lee P. Lim, Nelson C Lau, Earl G. Weinstein, Aliaa Abdelhakim, Soraya Yekta, Matthew W Rhoades, Christopher B. Burge, and David P. Bartel. The microRNAs of Caenorhabditis elegans. *Genes & development*, 17:991–1008, 2003. URL http://genesdev.cshlp.org/content/17/8/991.long.

[138] Lee P. Lim, Nelson C Lau, Philip Garrett-Engele, Andrew Grimson, Janell M Schelter, John Castle, David P. Bartel, Peter S Linsley, and Jason M Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–73, February 2005. ISSN 1476-4687. doi: 10.1038/nature03315. URL http://www.ncbi.nlm.nih.gov/pubmed/15685193.

[139] K Lindblad-Toh, CM Wade, TS Mikkelsen, and EK Karlsson. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438:803—-819, 2005. URL http://www.nature.com/nature/journal/v438/n7069/abs/nature04338.html.

[140] P.S. Linsley, Janell M Schelter, J. Burchard, M. Kibukawa, M.M. Martin, S.R. Bartz, J.M. Johnson, J.M. Cummins, C.K. Raymond, H. Dai, and Others. Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Molecular and Cellular Biology*, 27(6):2240, 2007. URL http://mcb.asm.org/cgi/content/abstract/27/6/2240.

[141] D Long, R Lee, P Williams, CY Chan, V Ambros, and Y Ding. Potent effect of target structure on microRNA function. *Nature structural & molecular biology*, 14:287—-294, 2007. URL http://www.nature.com/nsmb/journal/v14/n4/abs/nsmb1226.html.

[142] J Lu, G Getz, EA Miska, E Alvarez-Saavedra, J Lamb, D. Peck, A. Sweet-Cordero, B.L. Ebert, R.H. Mak, A.A. Ferrando, J.R. Downing, T. Jacks, H.R. Horvitz, and T.R. Golub. MicroRNA expression profiles classify human cancers. *Nature*, 435:834–838, 2005. URL http://www.nature.com/nature/journal/vaop/ncurrent/abs/nature03702.html.

[143] B M Lunde, C Moore, and G Varani. RNA-binding proteins: modular design for efficient function. *Nature reviews. Molecular cell biology*, 8:479–490, 2007.

[144] F.C. Lynn, P. Skewes-Cox, Y. Kosaka, M.T. McManus, B.D. Harfe, and M.S. German. MicroRNA Expression is Required for Pancreatic Islet Cell Genesis in the Mouse. *Diabetes*, pages 2938–2945, 2007.

[145] J R Lytle, T A Yario, and J A Steitz. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. *Proceedings of the National Academy of Sciences of the United States of America*, 104:9667–9672, 2007.

[146] William H Majoros and Uwe Ohler. Spatial preferences of microRNA targets in 3′ untranslated regions. *BMC genomics*, 8: 152, January 2007. ISSN 1471-2164. doi: 10.1186/1471-2164-8-152. URL http://www.ncbi.nlm.nih.gov/pubmed/17555584.

[147] K C Martin and A Ephrussi. mRNA Localization: Gene Expression in the Spatial Dimension. *Cell*, 136:719–730, 2009.

[148] Marbà Martina Maya. *Stochastic Modelling of Gene Regulation with microRNAs*. Master thesis, University of Glasgow, 2009.

[149] S Mayrand, B Setyono, J R Greenberg, and T Pederson. Structure of nuclear ribonucleoprotein: identification of proteins in contact with poly(A)+ heterogeneous nuclear RNA in living HeLa cells. *The Journal of Cell Biology*, 90:380–384, 1981.

[150] P. McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman & Hall / CRC, 1989.

[151] A E McKee, E Minet, C Stern, S Riahi, C D Stiles, and P A Silver. A genome-wide in situ hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain. *BMC Dev Biol.*, 5:14, 2005.

[152] K M Meisenheimer and T H Koch. Photocross-linking of nucleic acids to associated proteins. *Crit. Rev. Biochem. Mol. Biol.*, 32:101–140, 1997.

[153] Gunter Meister, Markus Landthaler, Agnieszka Patkaniowska, Yair Dorsett, Grace Teng, and Thomas Tuschl. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Molecular cell*, 15(2):185–97, July 2004. ISSN 1097-2765. doi: 10.1016/j.molcel.2004.07.007. URL http://www.ncbi.nlm.nih.gov/pubmed/15260970.

[154] Kevin C Miranda, Tien Huynh, Yvonne Tay, Yen-Sin Ang, Wai-Leong Tam, Andrew M Thomson, Bing Lim, and Isidore Rigoutsos. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, 126 (6):1203–17, 2006. ISSN 0092-8674. doi: 10.1016/j.cell.2006.07.031. URL http://dx.doi.org/10.1016/j.cell.2006.07.031.

[155] Eric A Miska, Ezequiel Alvarez-Saavedra, Matthew Townsend, Akira Yoshii, Nenad Sestan, Pasko Rakic, Martha Constantine-Paton, and H Robert Horvitz. Microarray analysis of microRNA expression in the developing mammalian brain. *Genome biology*, 5(9):R68, January 2004. ISSN 1465-6914. doi: 10.1186/gb-2004-5-9-r68. URL http://www.ncbi.nlm.nih.gov/pubmed/15345052.

[156] M J Moore and N J Proudfoot. Pre-mRNA Processing Reaches Back to Transcription and Ahead to Translation. *Cell*, 136:688–700, 2009.

[157] Zissimos Mourelatos, Josée Dostie, Sergey Paushkin, Anup Sharma, Bernard Charroux, Linda Abel, Juri Rappsilber, Matthias Mann, and Gideon Dreyfuss. miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes & development*, 16(6):720–8, 2002. ISSN 0890-9369. doi: 10.1101/gad.974702. URL http://genesdev.cshlp.org/cgi/content/abstract/16/6/720.

[158] Brian Munsky, Brooke Trinh, and Mustafa Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular systems biology*, 5(318):318, January 2009. ISSN 1744-4292. doi: 10.1038/msb.2009.75. URL http://www.ncbi.nlm.nih.gov/pubmed/19888213.

[159] Venugopal Nair and Mihaela Zavolan. Virus-encoded microRNAs: novel regulators of gene expression. *Trends in microbiology*, 14(4):169–75, 2006. ISSN 0966-842X. doi: 10.1016/j.tim.2006.02.007. URL http://dx.doi.org/10.1016/j.tim.2006.02.007.

[160] JA Nelder and R Mead. A simplex method for function minimization. *The Computer Journal*, 7:308–313, 1965. URL http://comjnl.oxfordjournals.org/cgi/content/abstract/7/4/308.

[161] Cydney B Nielsen, Noam Shomron, Rickard Sandberg, Eran Hornstein, Jacob Kitzman, and Christopher B Burge. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, 13(11):1894–910, November 2007. ISSN 1355-8382. doi: 10.1261/rna.768207. URL http://rnajournal.cshlp.org/cgi/content/abstract/13/11/1894.

[162] Ryan M O'Connell, Konstantin D Taganov, Mark P Boldin, Genhong Cheng, and David Baltimore. MicroRNA-155 is induced during the macrophage inflammatory response. *Proceedings of the National Academy of Sciences of the United States of America*, 104(5):1604–9, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0610731104. URL http://www.pnas.org/cgi/content/abstract/104/5/1604.

[163] Thomas Ohrt, Jörg Mütze, Wolfgang Staroske, Lasse Weinmann, Julia Höck, Karin Crell, Gunter Meister, and Petra Schwille. Fluorescence correlation spectroscopy and fluorescence cross-correlation spectroscopy reveal the cytoplasmic origination of loaded nuclear RISC in vivo in human cells. *Nucleic acids research*, 36(20):6439–49, November 2008. ISSN 1362-4962. doi: 10.1093/nar/gkn693. URL http://nar.oxfordjournals.org/cgi/content/abstract/36/20/6439.

[164] A J Olson, Julius Brennecke, A A Aravin, G J Hannon, and R Sachidanandam. Analysis of large-scale sequencing of small RNAs. In *Pacific Symposium on Biocomputing*, pages 126–36, January 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18229681.

[165] U A Orom, F C Nielsen, and A H Lund. MicroRNA-10a Binds the 5'UTR of Ribosomal Protein mRNAs and Enhances Their Translation. *Mol. Cell*, 30:460–471, 2008.

[166] Giorgos L Papadopoulos, Martin Reczko, Victor A Simossis, Praveen Sethupathy, and Artemis G Hatzigeorgiou. The database of experimentally supported targets: a functional update of TarBase. *Nucleic acids research*, 37(Database issue):D155–8, January 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn809. URL http://www.ncbi.nlm.nih.gov/pubmed/18957447.

[167] B J Park, J W Kang, S W Lee, S J Choi, Y K Shin, Y H Ahn, Y H Choi, D Choi, K S Lee, and S Kim. The haploinsufficient tumor suppressor p18 upregulates p53 via interactions with ATM/ATR. *Cell*, 120:209–221, 2005.

[168] J C Parker, K M Andrews, M R Allen, J L Stock, and J D McNeish. Glycemic control in mice with targeted disruption of the glucagon receptor gene. *Biochem Biophys Res Commun.*, 290:839–843, 2002.

[169] Sébastien Pfeffer, Mihaela Zavolan, Friedrich A Grässer, Minchen Chien, James J Russo, Jingyue Ju, Bino John, Anton J Enright, Debora Marks, Chris Sander, and Thomas Tuschl. Identification of virus-encoded microRNAs. *Science*, 304(5671):734–6, 2004. ISSN 1095-9203. doi: 10.1126/science.1096781. URL http://www.sciencemag.org/cgi/content/abstract/304/5671/734.

[170] Matthew N Poy, Lena Eliasson, Jan Krützfeldt, Satoru Kuwajima, Xiaosong Ma, Patrick E Macdonald, Sébastien Pfeffer, Thomas Tuschl, Nikolaus Rajewsky, Patrik Rorsman, and Markus Stoffel. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature*, 432(7014):226–30, November 2004. ISSN 1476-4687. doi: 10.1038/nature03076. URL http://www.ncbi.nlm.nih.gov/pubmed/15538371.

[171] M.N. Poy, Jean Hausser, Mirko Trajkovski, Matthias Braun, Stephan Collins, Patrik Rorsman, Mihaela Zavolan, and Markus Stoffel. miR-375 maintains normal pancreatic alpha-and beta-cell mass. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14):5813, 2009. URL http://www.pnas.org/content/106/14/5813.full.

[172] KD Pruitt, T Tatusova, and DR Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 33: D501–D504, 2005. URL http://nar.oxfordjournals.org/cgi/content/abstract/gkl842v1.

[173] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL http://www.r-project.org.

[174] Arjun Raj, Patrick Van Den Bogaard, Scott A Rifkin, Alexander Van Oudenaarden, and Sanjay Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, 5(10):877–879, 2008. doi: 10.1038/NMETH.1253.

[175] Nikolaus Rajewsky. microRNA target predictions in animals. *Nature genetics*, 38 Suppl:S8–13, June 2006. ISSN 1061-4036. doi: 10.1038/ng1798. URL http://www.ncbi.nlm.nih.gov/pubmed/16736023.

[176] Nikolaus Rajewsky and Nicholas D Socci. Computational identification of microRNA targets. *Developmental biology*, 267(2):529–35, March 2004. ISSN 0012-1606. doi: 10.1016/j.ydbio.2003.12.003. URL http://www.ncbi.nlm.nih.gov/pubmed/15013811.

[177] Christopher K Raymond, Brian S Roberts, Phillip Garrett-Engele, Lee P Lim, and Jason M Johnson. Simple, quantitative primer-extension PCR assay for direct monitoring of microRNAs and short-interfering RNAs. *RNA*, 11(11):1737–44, November 2005. ISSN 1355-8382. doi: 10.1261/rna.2148705. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1370860&tool=pmcentrez&rendertype=abstract.

[178] Brenda J. Reinhart, Frank J. Slack, Michael Basson, Amy E. Pasquinelli, Jill C. Bettinger, Ann E. Rougvie, H. Robert Horvitz, and Gary Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772): 901–906, 2000. URL http://www.nature.com/nature/journal/v403/n6772/abs/403901a0.html.

[179] J C Reyes, J Barra, C Muchardt, A Camus, C Babinet, and M Yaniv. Altered control of cellular proliferation in the absence of mammalian brahma (SNF2alpha). *EMBO J*, 17:6979–6991, 1998.

[180] Matthew W Rhoades, Brenda J Reinhart, Lee P. Lim, Christopher B Burge, Bonnie Bartel, and David P. Bartel. Prediction of plant microRNA targets. *Cell*, 110(4):513–20, August 2002. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/12202040.

[181] Harlan Robins and William H Press. Human microRNAs target a functionally distinct population of genes with AT-rich 3′ UTRs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15557–62, October 2005. ISSN 0027-8424. doi: 10.1073/pnas.0507443102. URL http://www.ncbi.nlm.nih.gov/pubmed/16230613.

[182] Harlan Robins, Ying Li, and Richard W Padgett. Incorporating structure to predict microRNA targets. *Proceedings of the National Academy of Sciences of the United States of America*, 102(11):4006–9, March 2005. ISSN 0027-8424. doi: 10.1073/pnas.0500775102. URL http://www.ncbi.nlm.nih.gov/pubmed/15738385.

[183] R G Roeder and W J Rutter. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature*, 224(5216): 234–7, October 1969. ISSN 0028-0836. URL http://www.ncbi.nlm.nih.gov/pubmed/5344598.

[184] JG Ruby, C Jan, C Player, MJ Axtell, W Lee, C Nusbaum, H Ge, and David P. Bartel. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, 127(6):1193–1207, 2006. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867406014681.

[185] Mark A Samols, Rebecca L Skalsky, Ann M Maldonado, Alberto Riva, M Cecilia Lopez, Henry V Baker, and Rolf Renne. Identification of cellular genes targeted by KSHV-encoded microRNAs. *PLoS pathogens*, 3(5):e65, May 2007. ISSN 1553-7374. doi: 10.1371/journal.ppat.0030065. URL http://www.ncbi.nlm.nih.gov/pubmed/17500590.

[186] J R Sanford, X Wang, M Mort, N Vanduyn, D N Cooper, S D Mooney, H J Edenberg, and Y Liu. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome research*, 19:381–394, 2009.

[187] Richa Saxena, Benjamin F Voight, Valeriya Lyssenko, Noël P Burtt, Paul I W de Bakker, Hong Chen, Jeffrey J Roix, Sekar Kathiresan, Joel N Hirschhorn, Mark J Daly, Thomas E Hughes, Leif Groop, David Altshuler, Peter Almgren, Jose C Florez, Joanne Meyer, Kristin Ardlie, Kristina Bengtsson Boström, Bo Isomaa, Guillaume Lettre, Ulf Lindblad, Helen N Lyon, Olle Melander, Christopher Newton-Cheh, Peter Nilsson, Marju Orho-Melander, Lennart Rå stam, Elizabeth K Speliotes, Marja-Riitta Taskinen, Tiinamaija Tuomi, Candace Guiducci, Anna Berglund, Joyce Carlson, Lauren Gianniny, Rachel Hackett, Liselotte Hall, Johan Holmkvist, Esa Laurila, Marketa Sjögren, Maria Sterner, Aarti Surti, Margareta Svensson, Malin Svensson, Ryan Tewhey, Brendan Blumenstiel, Melissa Parkin, Matthew Defelice, Rachel Barry, Wendy Brodeur, Jody Camarata, Nancy Chia, Mary Fava, John Gibbons, Bob Handsaker, Claire Healy, Kieu Nguyen, Casey Gates, Carrie Sougnez, Diane Gage, Marcia Nizzari, Stacey B Gabriel, Gung-Wei Chirn, Qicheng Ma, Hemang Parikh, Delwood Richardson, Darrell Ricke, and Shaun Purcell. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–6, June 2007. ISSN 1095-9203. doi: 10.1126/science.1142358. URL http://www.ncbi.nlm.nih.gov/pubmed/17463246.

[188] Thomas D Schmittgen, Eun Joo Lee, Jinmai Jiang, Anasuya Sarkar, Liuqing Yang, Terry S Elton, and Caifu Chen. Real-time PCR quantification of precursor and mature microRNA. *Methods*, 44(1):31–8, January 2008. ISSN 1046-2023. doi: 10.

1016/j.ymeth.2007.09.006. URL http://www.ncbi.nlm.nih.gov/pubmed/18158130.

[189] I Schomburg, A Chang, C Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucleic acids research*, 32:D431–433, 2004. URL http://nar.oxfordjournals.org/cgi/content/abstract/32/suppl_1/D431.

[190] Hervé Seitz. Redefining microRNA targets. *Current biology*, 19 (10):870–3, 2009. ISSN 1879-0445. doi: 10.1016/j.cub.2009.03.059. URL http://www.ncbi.nlm.nih.gov/pubmed/19375315.

[191] M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, 2008. URL http://www.nature.com/nature/journal/vaop/ncurrent/full/nature07228.html.

[192] Reut Shalgi, Daniel Lieber, Moshe Oren, and Yitzhak Pilpel. Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS computational biology*, 3(7):e131, 2007. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0030131. URL http://www.ncbi.nlm.nih.gov/pubmed/17630826.

[193] P M Sharp and W H Li. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*, 15:1281–1295, 1987.

[194] Rui Shi and Vincent Chiang. Facile means for quantifying microRNA expression by real-time PCR. *BioTechniques*, 39(4):519–525, October 2005. ISSN 0736-6205. doi: 10.2144/000112010. URL http://www.biotechniques.com/article/05394ST05.

[195] C Shih and R A Weinberg. Isolation of a transforming sequence from a human bladder carcinoma cell line. *Cell*, 29(1):161–9, May 1982. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/6286138.

[196] D.Q. Shih, M. Heimesaat, S. Kuwajima, R. Stein, C.V. Wright, and M. Stoffel. Profound defects in pancreatic beta-cell function in mice with combined heterozygous mutations in Pdx-1, Hnf-1alpha, and Hnf-3beta. *Proceedings of the National Academy of Sciences of the United States of America*, 99:3818–3823, 2002.

[197] R Siddharthan, E D Siggia, and E van Nimwegen. PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny. *PLoS computational biology*, 1:e67, 2005.

[198] Lasse Sinkkonen, Tabea Hugenschmidt, Philipp Berninger, Dimos Gaidatzis, Fabio Mohn, Caroline G. Artus-Revel, Mihaela Zavolan, Petr Svoboda, and Witold Filipowicz. MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. *Nature structural & molecular biology*, 15(3):259, 2008.

[199] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3:Article3, January 2004. ISSN 1544-6115. doi: 10.2202/1544-6115.1027. URL http://www.ncbi.nlm.nih.gov/pubmed/16646809.

[200] N Sonenberg and A G Hinnebusch. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell*, 136:731–745, 2009.

[201] KA Spackman. Signal detection theory: Valuable tools for evaluating inductive learning. In Morgan Kaufman, editor, *Sixth international workshop on Machine learning*, pages 160–163, 1989. URL http://portal.acm.org/citation.cfm?id=102172.

[202] Stephen a Stanhope, Srikumar Sengupta, Johan den Boon, Paul Ahlquist, and Michael a Newton. Statistical use of argonaute expression and RISC assembly in microRNA target identification. *PLoS computational biology*, 5(9):e1000516, 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000516. URL http://www.ncbi.nlm.nih.gov/pubmed/19779550.

[203] Alexander Stark, Julius Brennecke, Robert B Russell, and Stephen M Cohen. Identification of Drosophila MicroRNA targets. *PLoS biology*, 1(3):E60, December 2003. ISSN 1545-7885. doi: 10.1371/journal.pbio.0000060. URL http://www.ncbi.nlm.nih.gov/pubmed/14691535.

[204] Alexander Stark, Julius Brennecke, Natascha Bushati, Robert B Russel, and Stephen M Cohen. Animal microRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123:1133–1146, 2005.

[205] AI Su, T Wiltshire, S Batalov, H Lapp, KA Ching, D Block, J Zhang, R Soden, M Hayakawa, G Kreiman, MP Cooke, JR Walker, and JB Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16): 6062–6067, 2004. URL http://www.pnas.org/content/101/16/6062.abstract.

[206] Christopher S Sullivan, Adam T Grundhoff, Satvir Tevethia, James M Pipas, and Don Ganem. SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. *Nature*, 435(7042):682–6, June 2005. ISSN 1476-4687. doi: 10.1038/nature03576. URL http://www.ncbi.nlm.nih.gov/pubmed/15931223.

[207] Hakim Tafer, Stefan Ludwig Ameres, Gregor Obernosterer, Christoph A. Gebeshuber, Renée Schroeder, Javier Martinez, and Ivo L Hofacker. The impact of target site accessibility on the design of effective siRNAs. *Nature biotechnology*, 26(5):578–83, May 2008. ISSN 1546-1696. doi: 10.1038/nbt1404. URL http://www.ncbi.nlm.nih.gov/pubmed/18438400.

[208] Yvonne Tay, Jinqiu Zhang, Andrew M Thomson, Bing Lim, and Isidore Rigoutsos. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, 455(7216):1124–8, October 2008. ISSN 1476-4687. doi: 10.

1038/nature07299. URL http://www.ncbi.nlm.nih.gov/pubmed/18806776.

[209] S A Tenenbaum, C C Carson, P J Lager, and J D Keene. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 97:14085–14090, 2000.

[210] To-Ha Thai, Dinis Pedro Calado, Stefano Casola, K Mark Ansel, Changchun Xiao, Yingzi Xue, Andrew Murphy, David Frendewey, David Valenzuela, Jeffery L Kutok, Marc Schmidt-Supprian, Nikolaus Rajewsky, George Yancopoulos, Anjana Rao, and Klaus Rajewsky. Regulation of the germinal center response by microRNA-155. *Science*, 316(5824):604–8, 2007. ISSN 1095-9203. doi: 10.1126/science.1141229. URL http://www.sciencemag.org/cgi/content/abstract/316/5824/604.

[211] Rolf Thermann and Matthias W Hentze. Drosophila miR2 induces pseudo-polysomes and inhibits translation initiation. *Nature*, 447(7146):875–8, June 2007. ISSN 1476-4687. doi: 10.1038/nature05878. URL http://www.ncbi.nlm.nih.gov/pubmed/17507927.

[212] Shogo Tokumaru, Motoshi Suzuki, Hideki Yamada, Masato Nagino, and Takashi Takahashi. let-7 regulates Dicer expression and constitutes a negative feedback loop. *Carcinogenesis*, 29(11):2073–7, November 2008. ISSN 1460-2180. doi: 10.1093/carcin/bgn187. URL http://www.ncbi.nlm.nih.gov/pubmed/18700235.

[213] Susan Tweedie, Michael Ashburner, Kathleen Falls, Paul Leyland, Peter McQuilton, Steven Marygold, Gillian Millburn, David Osumi-Sutherland, Andrew Schroeder, Ruth Seal, and Haiyan Zhang. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic acids research*, 37(Database issue):D555–9, January 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn788. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686450&tool=pmcentrez&rendertype=abstract.

[214] J Ule, K B Jensen, M Ruggiu, A Mele, A Ule, and R B Darnell. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302:1212–1215, 2003.

[215] Jernej Ule, Giovanni Stefani, Aldo Mele, Matteo Ruggiu, Xuning Wang, Bahar Taneri, Terry Gaasterland, Benjamin J Blencowe, and Robert B Darnell. An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444(7119):580–6, November 2006. ISSN 1476-4687. doi: 10.1038/nature05304. URL http://www.ncbi.nlm.nih.gov/pubmed/17065982.

[216] G Vaidyanathan, M J Cismowski, G Wang, T S Vincent, K D Brown, and S M Lanier. The Ras-related protein AGS1/RASD1 suppresses cell growth. *Oncogene*, 23:5858–5863, 2004.

[217] Roberto Valverde, Laura Edwards, and Lynne Regan. Structure and function of KH domains. *The FEBS journal*, 275(11):2712–26, June 2008. ISSN 1742-464X. doi: 10.1111/j.1742-4658.2008.06411.x. URL http://www.ncbi.nlm.nih.gov/pubmed/18422648.

[218] Erik van Nimwegen. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC bioinformatics*, 8 Suppl 6:S4, January 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-S6-S4. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1995539&tool=pmcentrez&rendertype=abstract.

[219] E. van Rooij, L.B. Sutherland, X. Qi, J.A. Richardson, J. Hill, and E.N. Olson. Control of Stress-Dependent Cardiac Growth and Gene Expression by a MicroRNA. *Science*, 316:575–579, 2007.

[220] S. Vasudevan, Y. Tong, and J.A. Steitz. Switching from repression to activation: microRNAs can up-regulate translation. *Science*, 318(5858):1931, 2007. URL http://www.sciencemag.org/cgi/content/abstract/sci;318/5858/1931.

[221] M C Vella, E Y Choi, S Y Lin, K Reinert, and F J Slack. The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. *Genes & development*, 18:132–137, 2004.

[222] A J Wagenmakers, R J Reinders, and W J van Venrooij. Cross-linking of mRNA to proteins by irradiation of intact cells with ultraviolet light. *Eur. J. Biochem.*, 112:323–330, 1980.

[223] S Wang, A B Aurora, B A Johnson, X Qi, J McAnally, J A Hill, J A Richardson, R Bassel-Duby, and E N Olson. The endothelial-specific microRNA miR-126 governs vascular integrity and angiogenesis. *Dev Cell*, 15:261–271.

[224] X Wang, J McLachlan, P D Zamore, and T M T Hall. Modular Recognition of RNA by a Human Pumilio-Homology Domain. *Cell*, 110:501–512, 2002.

[225] Xia Wang, Yan Li, Xue Xu, and Yong-hua Wang. Toward a system-level understanding of microRNA pathway via mathematical modeling. *Bio Systems*, 100(1):31–8, April 2010. ISSN 1872-8324. doi: 10.1016/j.biosystems.2009.12.005. URL http://www.ncbi.nlm.nih.gov/pubmed/20005918.

[226] Xiaowei Wang and Xiaohui Wang. Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic acids research*, 34(5):1646–52, March 2006. ISSN 1362-4962. doi: 10.1093/nar/gkl068. URL http://nar.oxfordjournals.org/cgi/content/abstract/34/5/1646.

[227] Yanli Wang, Stefan Juranek, Haitao Li, Gang Sheng, Thomas Tuschl, and Dinshaw J. Patel. Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature*, 456(18):921, 2008.

[228] Yanli Wang, Stefan Juranek, Haitao Li, Gang Sheng, Greg S Wardle, Thomas Tuschl, and Dinshaw J Patel. Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. *Nature*, 461(7265):754–61, 2009. ISSN 1476-4687. doi: 10.1038/nature08434. URL http://www.ncbi.nlm.nih.gov/pubmed/19812667.

[229] RH Waterston, K Lindblad-Toh, E Birney, and J Rogers. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

[230] M Wickens, D S Bernstein, J Kimble, and R Parker. A PUF family portrait: 3'UTR regulation as a way of life. *Trends Genet.*, 18:150–157, 2002.

[231] B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75:855–862, 1993.

[232] Thomas D Wu and Colin K Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics (Oxford, England)*, 21(9):1859–75, May 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti310. URL http://www.ncbi.nlm.nih.gov/pubmed/15728110.

[233] Zhijn Wu, R.A. Irizarry, Robert Gentleman, Francisco Martinez-Murillo, and F. Spencer. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 99(468):909–917, 2004. URL http://pubs.amstat.org/doi/abs/10.1198/016214504000000683.

[234] Edward Yang, Erik van Nimwegen, Mihaela Zavolan, Nikolaus Rajewsky, Mark Schroeder, Marcelo Magnasco, and James E Darnell. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome research*, 13(8): 1863–72, August 2003. ISSN 1088-9051. doi: 10.1101/gr.1272403. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=403777&tool=pmcentrez&rendertype=abstract.

[235] G W Yeo, N G Coufal, T Y Liang, G E Peng, X.-D. Fu, and F H Gage. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature structural & molecular biology*, 16:130–137, 2009.

[236] R. Yi, D. O'Carroll, H.A. Pasolli, Z. Zhang, F.S. Dietrich, A. Tarakhovsky, and E. Fuchs. Morphogenesis in skin is governed by discrete sets of differentially expressed microRNAs. *Nature Genetics*, 38:356–362, 2006.

[237] J K Yisraeli. VICKZ proteins: a multi-talented family of regulatory RNA-binding proteins. Biol. *Cell*, 97:87–96, 2005.

[238] Yong Zhao, Joshua F Ransom, Ankang Li, Vasanth Vedantham, Morgan von Drehle, Alecia N Muth, Takatoshi Tsuchihashi, Michael T McManus, Robert J Schwartz, and Deepak Srivastava. Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. *Cell*, 129(2):303–17, April 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.03.030. URL http://www.ncbi.nlm.nih.gov/pubmed/17397913.

[239] V.P. Zhdanov. Conditions of appreciable influence of microRNA on a large number of target mRNAs. *Molecular BioSystems*, 5(6):638–643, 2009. URL http://www.rsc.org/delivery/_ArticleLinking/ArticleLinking.asp?JournalCode=MB&Year=2009&ManuscriptID=b808095j&Iss=6.

[240] Jakob T Zipprich, Sankar Bhattacharyya, Hansruedi Mathys, and Witold Filipowicz. Importance of the C-terminal domain of the human GW182 protein TNRC6C for translational repression. *RNA*, 15(5):781–93, 2009. ISSN 1469-9001. doi: 10.1261/rna.1448009. URL http://rnajournal.cshlp.org/cgi/content/abstract/15/5/781.

[241] Dimitrios G Zisoulis, Michael T Lovci, Melissa L Wilbert, Kasey R Hutt, Tiffany Y Liang, Amy E Pasquinelli, and Gene W Yeo. Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nature structural & molecular biology*, 17(2):173–9, February 2010. ISSN 1545-9985. doi: 10.1038/nsmb.1745. URL http://dx.doi.org/10.1038/nsmb.1745.