CHARACTERIZATION OF POST-TRANSCRIPTIONAL REGULATORY NETWORK OF RNA-BINDING PROTEINS USING COMPUTATIONAL PREDICTIONS AND DEEP SEQUENCING DATA

Inauguraldissertation

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

von

mohsen khorshid

aus dem Iran

Basel, 2013

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Mihaela Zavolan & Prof. Dr. Sven Bergmann

Basel, den 21 Februar 2012

Prof. Dr. Martin Spiess
Dekan

Mohsen Khorshid: *Characterization of post-transcriptional regulatory network of RNA-binding proteins using computational predictions and deep sequencing data*, PhD Thesis , 2013

Family means nobody gets left behind, or forgotten.

— *Lilo & Stitch*

Dedicated to the loving memory of my mother

1947 – 2011

Dedicated also to my beloved parents, brothers, wife and all my family members

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

1. **Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP**
   Authors: Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Jr. Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan and Thomas Tuschl

   (2010) *Hafner et al.Cell, doi:10.1016/j.cell.2010.03.009*

2. **CLIPZ: A database and analysis environment for experimentally-determined binding sites of RNA-binding proteins**
   Authors: Mohsen Khorshid, Christoph Rodak and Mihaela Zavolan

   (2010) *Khorshid et al.Nucleic Acids Research, doi: 10.1093/nar/gkq940*

3. **A quantitative analysis of CLIP methods for identifying binding sites of RNRNA-binding proteins**
   Authors: Shivendra Kishore, Lukasz Jaskiewicz, Lukas Burger, Jean Hausser, Mohsen Khorshid and Mihaela Zavolan

   (2011) *Kishore et al.Nature Methods, doi:10.1038/nmeth.1608*

4. **A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets**
   Authors: Mohsen Khorshid, Jean Hausser, Mihaela Zavolan, and Erik van Nimwegen.

   (2013) *Khorshid et al.Nature Methods , 12:46 doi:10.1038/nmeth.2341*

ACKNOWLEDGMENTS

The successful completion of this thesis would not have been possible without the assistance and support of many people.

First of all, I would like to express my gratitude to my mother and father, who always encouraged me by all means and giving me all possible support I could think of. I am very grateful to my best friend and wife Aida who have given me the freedom to make my own decisions. The same goes to my brothers, Arash, Babak, Saeed, Mina, Shadi and all my family members.

I would like to especially thank Mihaela Zavolan offered me the opportunity to undertake the work towards this degree. I always found it quite interesting to answer challenging questions and it could not have been possible without her support and advice throughout the years. I am especially indebted to Peter R. Wills for his guidance toward finding this opportunity at first place.

I am very grateful to Erik van Nimwegen and Thomas Tuschl for a textbook example of collaborations, useful discussions, and fruitful entertaining meetings.

I am thankful to Yvonne Steger, Rita Manohar and Sarah Güthe for taking care of all the administrative things during these last 4 years, thereby making my life so much easier. Also, special thanks to the IT admins, Konstantin Arnold, Rainer, and Jan for taking care of all the mess I made.

It was a great time working with Christoph Rodak and I am thankful for it. I would like to thank Anneke Brümmer, Andreas Gruber, Jean Hausser, Lukasz Jaskiewicz, Shivendra Kishore, Nitish Mittal, Biter Bilen, Yoana Aleksandrova Dimitrova, the other Andreas Gruber, Dominik Jedlinski, Hadi Jorjani, Afzal Pasha Syed, Philipp Berninger, Piotr Balwierz, Lukas Burger, Manuel Ascano, Markus Hafner and Georges Martin for invaluable help with many problems I encountered in my work.

Also, I am much obligated to Shivendra Kishore, Nicholas William Kelley, Olin Silander, Andreas Gruber and Peter Pemberton-Ross for proofreading my manuscripts and eliminating lots of mistakes.

And finally, I would like to express my gratitude to Mihaela Zavolan, Erik van Nimwegen, Sven Bergman, Niko Beerenwinkel, and Markus Rüegg for serving on my examining committee.

To everyone I mentioned and those I forgot: **Thank you very much!**

CONTENTS

LIST OF FIGURES

Part I

INTRODUCTION

1 INTRODUCTION

Life is not measured by the number of breaths we take but by the moments that take our breath away.

–Hilary Cooper

1.1 BACKGROUND

Life is complex. The definition of life has been debated among many scientists and philosophers. In 1944, Erwin Schrödinger[1] in his famous article, *What is life?*, stated that life is not a closed system. This is simply because a world governed by the second law of thermodynamics [154] has a tendency to achieve a state of maximum disorder. However, life approaches and maintains a highly ordered system.

The ability of organisms to maintain order in a world governed by the second law of thermodynamics has to do with context and hierarchy. The phenomenon of *heredity* plays a fundamental role in this process. Schrödinger anticipated that something like DNA exists and because of that order is maintained from parent to progeny and *"genes"* carry the hypothetical material of a *"definite hereditary feature"*. Material and energy are inherited from one generation to another. That explains why we do not get something from nothing.

Organisms' DNA codes for all the RNA and protein molecules required to construct its cells. The cell types in a multicellular organism differentiate from other cell types based on the synthesis and accumulation of different sets of RNA, protein, lipids and carbohydrate biomolecules. Based on this mechanism, much phenotypical diversity can be derived [1].

The production of an observable molecular product (e.g. RNA or protein) by a gene is defined as *gene expression* [1]. In general, gene expression that underlies the development of multicellular organisms does not rely on changes in the DNA sequences of the corresponding gene.

There are, however, a few cases where DNA rearrangements of the genome take place during the development of an organism. Perhaps the most impressive examples of programmed DNA rearrangement take place in bone-marrow-derived (B) cells and thymus-derived (T) cells that play a role the immune system of mammals [70].

1.1.1 *Regulation of gene expression*

Gene expression is regulated at multiple levels including:

[1] The Nobel Prize in Physics 1933 was awarded jointly to Erwin Schrödinger & Paul Adrien Maurice Dirac for the discovery of new productive forms of atomic theory http://www.nobelprize.org/nobel_prizes/physics/laureates/1933/

- **Transcriptional control** that regulates the timing and the level of transcription for a given gene [1].
- **RNA processing control** that regulates splicing, 3′ end formation, RNA editing and processing of RNA [1, 129, 144].
- **RNA transport control** that selects which processed RNAs are exported from the nucleus to cytosol [81].
- **RNA localization control** that determines where to keep the transported transcripts.
- **mRNA degradation control** that destabilizes certain mRNA molecules in the cytoplasm [189].
- **Translational control** that decides which mRNAs in the cytoplasm are translated by ribosomes.
- **Protein folding** whereby function of protein is established through production of the correct structure.
- **Protein activity control** that activates, inactivates, localizes or transports specific protein molecules after they have been made [1].

1.1.2 *Post-transcriptional regulation of gene expression*

In general, any mechanism that controls the gene expression at the level of RNA is part of the so-called *"post-transcriptional regulation"* of gene expression [1, 189]. Especially in *eukaryotes*, RNA found in the nucleus is more complex than that found in the cytoplasm: more than 95% of the RNA bases synthesized by RNA polymerase II never reach the cytoplasm. The main reason for this is the removal of introns, which account for 80% of the total bases [80]. This process is called *RNA splicing*.

Another example related to the post-transcriptional regulation of gene expression, is the study [155] by Schwanekamp et al. in 2006 to find out how extensively genes are regulated by post-transcriptional regulation. They monitored the effect of *dioxin*[2]. AHR gets activated by dioxin and include a set of genes encoding xenobiotic metabolizing proteins in order to enhance the body's main molecular defence against environmental toxins. Schwanekamp et al. wanted to find out whether toxicants such as dioxin significantly affect nuclear RNA levels and that cytoplasmic RNA levels are dependent on nuclear RNA levels. They compared nuclear and cytoplasmic RNA levels from untreated and dioxin-treated mouse embryonic fibroblasts. The result showed that nuclear RNA levels are strongly affected by dioxin due to effects

[2] Pervasive teratogen and carcinogen 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD or dioxin), One of the polycyclic aromatic hydrocarbon toxicants on Aryl-Hydrocarbon Receptor (AHR)

of proteins involved in nuclear RNA processing and transcription mostly affected the nuclear RNAs. The correlation between nuclear and cytoplasmic RNA levels is weak suggesting other regulatory mechanisms which control cytoplasmic RNA levels. AHR regulates key xenobiotic metabolizing genes at the transcriptional level, a larger impact of the dioxin-activated AHR are at post-transcriptional levels.

1.1.3 *Role of RNA-binding proteins in post-transcriptional regulation of gene expression*

RNA-binding proteins regulate the post-transcriptional events [67, 94, 52], such as RNA splicing [1, 129, 144] and editing [127] and also translation of RNA [8]. RNA binding proteins post-transcriptionally regulate a large amount of the transcriptome. This makes them interesting to the scientific community. *RNA interference* (RNAi) and *microRNAs* are both examples of post transcriptional regulation [10, 23, 67, 94, 134], which regulate the degradation of RNA and change the chromatin structure, see also 1.1.4.2.

Computational models describing the binding specificity of RBPs are lacking. This is in contrast, for instance, with transcription factors. Similarly, binding specificities of transcription factors have been catalogued in databases such as TRANSFAC[3] [130], but such databases are not common for RNA-binding proteins. Zheng et al.[200] designed a knowledge-based resource to predict the specificity and relative binding energy of RNA-binding proteins. However structural studies [7] suggest that the specificity of RNA-binding proteins may come from their multi-domain structure, each of the domains engaging only a few nucleotides.

Precise knowledge of the spatio-temporal associations between RBPs and mRNAs under various conditions is crucial to understand how the level, translation rate and cellular localization of those mRNAs are regulated during the life time of a cell. It is therefore clear that we need to first determine which RNA binding proteins associate with each mRNA and under which conditions. At the same time, such information could assist in defining the binding specificity of the protein of interest.

1.1.3.1 *Target site identification of RNA-binding proteins*

In order to study post transcriptional regulation several techniques are used. One of the initial methods to experimentally determine the set of targets for RNA binding proteins is the use of Differential Display Assay reverse transcription PCR (DDRT-PCR) [119]. It is a PCR-based method that allows extensive analysis of gene expression among several cell populations [159]. It selectively amplifies large numbers

[3] http://www.gene-regulation.com/pub/databases.html

of expressed sequences in an individual analysis, and then *"displays"* the genes by gel electrophoresis. In 2000, Sturtevant [167] reviewed the application of this method and addressed various limitations, including the large number of false-positive results and the difficulty in confirming differential expression.

Other methods are based on immunoprecipitation(IP)[4] of associated RNAs. RNA immunoprecipitation followed by microarray chip (RIP-Chip) can be used to detect the association of individual proteins with specific RNAs [86, 157]. Briefly, a subset of total cell mRNAs associated to endogenous mRNA-protein (mRNP) complexes is directly isolated and later identified using cDNA microarrays. First, the cells are harvested e.g. by treatment of cells with formaldehyde to cross-link in vivo Protein-RNA complexes. The next step is to conduct nuclei isolation and nuclear pellets lysis followed up by shearing of chromatin. The endogenous mRNA-protein (mRNP) complexes (RNA binding protein (RBP) of interest together with the bound RNA) is purified using immunoprecipitation and unbound material is washed off. RNA that is bound to immunoprecipitated RBP is then purified. Next, Reverse transcription (RT) of RNA to cDNA is performed. Finally, if target is known qPCR is performed. When the target is not known cDNA libraries are created and microarrays and sequencing can be used for target identification analysis.

Ule et al.[176] introduced a protocol in which UV crosslinking is used to isolate the binding sites of a particular RNA binding protein. The RNA is fragmented using *ribonucleases* (RNases) to get the RNA, which ideally contains the binding site extended by possible short flanking nucleotides. Next, the protein and the associated RNA fragments are isolated by immunoprecipitation, the protein is digested and the remaining RNA fragments are sequenced, currently by using a high-throughput sequencing machine. A few variants, known as HITS-CLIP [120, 25], PAR-CLIP [67], modified PAR-CLIP[95] and iCLIP [96], have been proposed. This method is referred to as *Crosslinking and immunoprecipitation* (CLIP) the characteristics of these methods is discussed in details in this thesis.

1.1.4 *Role of RNA interference and microRNAs in post-transcriptional regulation of gene expression*

In this sections the role of RNAi and microRNA regulation as examples of post transcriptional regulation of gene expression is elaborated more in details. In 1998, Andrew Z. Fire and Craig C. Mello[5] observed effective silencing of genes based on sequence specificity when they

[4] Immunoprecipitation (IP) is an antibody-based technique of precipitating a protein antigen out of solution using an antibody that specifically binds to that particular protein

[5] The Nobel Prize in Physiology or Medicine 2006 was awarded jointly to Andrew Z. Fire & Craig C. Mello for their discovery of RNA interference - gene silencing

exposed *Caenohabtidis elegans* to double-stranded RNAs (dsRNAs) [41]. Similar results were observed in *Drosophila* embryos [88]. Those observations created an exciting new field in RNA biology. Fire & Mello defined RNAi as an evolutionary-conserved gene-silencing mechanism that uses short, double-stranded RNAs (dsRNAs) to identify complementary target RNAs for sequence-specific degradation [41].

RNAi offers a powerful tool to specifically direct the degradation of complementary RNAs, and thus has great therapeutic potential in targeting diseases[6] [140]. The presence of RNAs of about 22 nucleotides in length[68, 69] that are complementary to the gene that is being suppressed is essential for RNAi. Despite our knowledge of the mechanism of RNAi, there is still a need for new techniques that will allow for a detailed mechanistic characterization of RNA-induced silencing complex (RISC) assembly and activity to further improve the biocompatibility of modified siRNAs [69, 140].

In 1993, Lee et al. identified the first microRNA [109]. In this study, Victor Ambros and his team positionally cloned the *lin-4* gene in worm and realized that no protein is encoded by this gene. *Lin-4* is a locus required for the correct timing of development in *Caenorhabditis elegans*. The interesting finding was that *lin-4* instead encompasses two *small noncoding RNAs* (ncRNAs), one 22 nucleotides long, and a longer form (*lin-4L*). These ncRNAs fold into a hairpin structure. In 2000, Ruvkun and colleagues discovered that *let-7*, which also regulates developmental timing in worms, also encompasses for a noncoding RNA [149]. Because *lin-4* and *let-7* control developmental timing, they were referred to as small temporal RNAs (stRNAs). Later on, researchers were able to clone additionally some hundreds of 19-25 nucleotide stRNA-like RNAs from worms, flies, and human cells that are similar to stRNAs and which derived from longer stem-loop precursor RNAs[99, 108]. Thus, the longer *lin-4L* was called the precursor molecule of the mature *lin-4*. Landgraf et al.largely studied the expression patterns of the microRNAs between cell lineages and tissues [104]. Results of this study showed that these precursors are expressed in many different ways. Some are ubiquitously produced in large quantities, whereas many others are temporally regulated or expressed only in specific tissues [104]. Some microRNAs appear to be transcribed in "coordination regulation operons"[7] indicating that they are closely distributed in the genome and cleaved from their

by double-stranded RNA. http://www.nobelprize.org/nobel_prizes/medicine/laureates/2006/

[6] For example, Alnylum Pharmaceuticals http://www.alnylam.com is developing RNAi therapeutic for the treatment of hemophilia and rare bleeding disorders

[7] In 2002, Keene and Tenenbaum [85] defined post-transcriptional operons as clusters of genes physically ordered in the genome in a manner enabling them to be regulated as groups. Operons represent a powerful mechanism to organize and express genetic information as functionally related combinations of monocistronic mRNA.

stem-loop precursors from within a long, common transcript which referred to as microRNA clusters [100, 101, 174, 141].

Grishok et al. [62] found that worms that accumulate *lin-4/let-7* lack endoribonuclease coded by *alg-2*. These worms failed to form the germ line early in their development [27] which leads to a defunct RNAi mechanism. *ALG* is a homologue of the human *Argonaute*. Both of them are member of RNA-Induced Silencing Complex (RISC) family of proteins. They function not only in microRNA maturation but are also required in animals, plants, and fungi for a variety of RNA-silencing phenomena, including RNAi and co-suppression.

1.1.4.1 *microRNA biogenesis*

During the last decade, substantial efforts have been made toward uncovering the biogenesis of microRNAs, their molecular mechanisms and functional roles in a variety of cellular contexts. Lee et al. [111] showed that microRNAs are commonly transcribed by RNA polymerase II from intragenic and intergenic chromosomal DNA regions into long primary transcripts of various lengths (usually 1-3 kb), named *primary microRNAs* (pri-microRNAs). The RNAse complex composed of RNase III Drosha and DGCR8[8] endonucleolytically process the pri-microRNA. They produce a 70-100 nucleotide long hairpin-precursor structure [110, 105]. The processed pri-microRNA, called as precursor microRNA (pre-microRNA) is then transported to the cytoplasm by an exportin-5 dependent mechanism [10]. In 2004, Lee et al. [112] showed that once exported into the cytosol, the double-stranded pre-microRNAs is further cleaved by Dicer into a mature double-stranded microRNA of variable length (approximately 20-25 nucleotides). Cleavage results in an imperfect duplex that is unwound, and the strand with the weakest base pairing at the 5′ end guide strand is preferentially loaded into an Argonaute protein family. Therefore, the guide strand or mature microRNA is loaded into a RISC, while the passenger strand, also known as microRNA* is generally destroyed [78]. As of the writing of this report, 1527 human microRNA genes[9] have been identified and listed in the official microRNA database (miRBase) [60].

1.1.4.2 *microRNA regulatory function*

microRNAs regulate many fundamental biological processes [99]. In 2004, Poy et al. [145] showed that *Myotrophin* is a target of miR-375, suggesting that insulin secretion and exocytosis is regulated by miR-375. One year later, Krützfeldt et al. found another example of microRNA regulatory function. They showed that miR-122 regulates

[8] DGCR8 or *DiGeorge syndrome critical region gene 8*, acts as Drosha's cofactor. It is the double-stranded RNA-binding protein

[9] The miRBase Sequence Database – Release 18, See http://www.mirbase.org/

cholesterol biosynthesis genes. They first reduced the level of miR-122 in mice by administrating *antagomirs*[10], and then compared the expression levels of the affected genes relative to their expression levels in the control samples[11].

It has been shown that the microRNAs regulate many other fundamental biological processes. For example, expression of the miR302/367 cluster leads to potent and rapid reprogramming of mouse and human somatic cells to an induced pluripotent stem cell (iPSC[12]) state. In fact, the reprogramming process event does not require exogenous transcription factors [5]. Several recent studies discuss functional roles of microRNAs in cancer. For example, miR-200 family (miR-200a, miR-200b, miR-200c, miR-141 and miR-429) of microRNAs and to miR-205 [58, 59, 143] are shown to inhibit the epithelial-mesenchymal transition (EMT) programme and function as tumor suppressor. Other links to cancer are oncogenes reported to be targeted by microRNAs [20, 74, 138]. Moreover, microRNAs have been shown to regulate DNA methylation [161, 34], embryonic development [54] and immunity [173, 139].

1.1.4.3 *microRNA target site predictions*

Biochemical and structural studies of the RISC complex bound to target RNA in *Bacteria*[13] postulate to some extent a complex protein interaction between the microRNA and the Argonaute protein as well as an interaction between microRNA and its mRNA target binding site [186, 187]. microRNAs are bound by Argonaute (Ago/EIF2C) proteins causing translational inhibition and mRNA destabilization or inhibition of translation of partially-complementary target mRNAs [10].

In plants, microRNAs generally find their mRNA targets by extensive complementarity [10]. Predictions based on this assumption are highly reliable [150]. However, it is very rare to see such extensive complementarity with consequent cleavage of the targeted message in animals [28, 197]. This makes it challenging to develop a computational algorithm that predicts most of the regulatory targets on a genome-wide scale without producing too many false predictions [10].

[10] An *antagomir* is a small chemically modified, cholesterol-conjugated single-stranded oligonucleotide that is perfectly complementary to the specific microRNA. In order to make the antagomirs more resistant to degradation machinery, they usually have modifications, such as 2′-O-methyl (2′-OMe) and phosphothioates [98] groups. Beal et al. showed that 2′-OMe group at the editing site substantially reduces the deaminiation rate. It might have either mispairing at the cleavage site of *Argonaute-2* (Ago2) or base modification to inhibit Ago2 cleavage and microRNA activity. It appears that this inhibition is due to irreversible binding of the microRNA but even that is still not completely known.

[11] mock transfected

[12] iPSC cells, exhibit the morphology and growth properties of embryonic stem(ES) cells and express ES cell marker genes, See Takahashi and Yamanaka.

[13] The study was performed in gram negative eubacterium *Thermus thermophilus*

A breakthrough in terms of more accurate predictive models was the use of preferential evolutionary conservation [115, 44]. Based on evolutionary conservation, microRNA prection algorithms like TargetScan[14] [61] or ElMMO[15] [46] have the ability to distinguish microRNA target sites from the multitude of 3′ UTR segments that otherwise would score equally well with regard to the quality of microRNA pairing.

Important features for target site recognition include pairing to the target mRNA with the 6-8 nucleotides from the 5′ end of the microRNA (seed region). It was shown that seed pairing is not always sufficient for repression. For example, Ameres et al. [3] studied target complementarity to microRNAs in *Drosophila*. They found that the targets match only to a microRNA seed region did not get tailed and trimmed. In contrast, only when seed pairing is accompanied by extensive central and 3′ pairing [16] between the microRNA and the target then potent tailing and trimming was achieved[3].

The results from different target predictions for microRNAs are not compatible with each other, meaning that different approaches lead to very different sets of predicted targets. Based on computational predictions, it is estimated that many of the protein-coding genes in mammalian are regulated by microRNAs [40], and it is estimated that 10s to 100s of mRNAs are targeted by microRNA [116, 135].

1.2 INTRODUCTION TO THE CHAPTERS

In section 1.1.3, it was aimed to briefly introduce post-transcriptional regulation of gene expression. Identification of RBP and *micro-Ribonucleo-Protein complexes* (miRNP) interactions with the target RNA is critical, because it may lead to the discovery of specific combination of sites (or modules) that may control distinct cellular processes and pathways [67]. CLIP demonstrates that a transcript will generally be bound and regulated by multiple RBPs and miRNPs, the spatio-temporal and/or combination of which will determine the final gene-specific regulatory outcome.

This report is divided into three parts: *Data Analysis*, *Mathematical Modeling* and *Conclusion and future directions*. In the *Data Analysis* part, various methods and tools for characterizing the post-transcriptional regulatory networks of RNA-binding proteins are discussed and applied. Chapter 2 introduces PAR-CLIP, a method for transcriptome-wide identification of RNA binding proteins at nucleotide resolution. PAR-CLIP was successfully applied on RNA binding proteins and their binding specificity was characterized.

[14] http://www.targetscan.org/

[15] http://www.mirz.unibas.ch/ElMMo3/

[16] With eight or fewer mismatches with the 3′ end of the microRNA

Partly due to their vast volume, the data that were so far generated in CLIP experiments have not been put in a form that enables fast and interactive exploration of binding sites. To address this need, Chapter 3 presents CLIPZ[17], which is a database and analysis environment for various kinds of deep sequencing (and in particular CLIP) data, that aims to provide an open-access repository of information for post-transcriptional regulatory elements.

Chapter 4 revisits various CLIP methods. A set of ideas in terms of both experimental protocols and data analysis are presented to improve the quality and reproducibility of such experiments. In general, cytoplasmic RNAs are isolated in CLIP experiments. Like many high-throughput experiments, CLIP has a certain amount of isolated RNAs which do not represent regulatory binding sites. To improve the quality of the obtained RNAs, a set of novel methods for data analysis are also suggested. These methods are added as new tools to the CLIPZ analysis platform.

Argonaute CLIP data could in principle be beneficial in improving the microRNA target site predictions. However, several questions still remain which cannot be addressed using CLIP methods. For example:

- Argonaute CLIP data by default does not reveal which microRNAs are more likely to interact to the mRNA binding site at the time of cross-linking.

- As mentioned earlier, biochemical and structural studies of *Thermus thermophilus* Argonaute protein [186, 187] suggest that the protein-RNA interaction between microRNA and the Argonaute protein forms a physical structure that only some positions in the microRNA become accessible to the target binding site. Having inferred the interacting microRNA, it is also interesting to predict the most plausible secondary structure of the hybridized microRNA-mRNA complex.

Mathematical Modeling part of the report contains Chapter 5. This chapter presents a novel mathematical model called MIRZA[18] to address the above mentioned questions. An in-depth introduction to MIRZA is presented and its performance in terms of identifying functionally relevant targets of microRNAs is discussed.

Finally, *Conclusion and future directions* part of the report contains Chapter 6 in which discusses the main findings of the projects and gives an outlook of where future work could be taken up.

[17] http://www.clipz.unibas.ch

[18] Source code available at: http://www.mirz.unibas.ch/software.php

Part II

DATA ANALYSIS

2

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

The method presented in this chapter was developed in collaboration with the Tuschl lab at Rockefeller University, New York and were originally published in Cell [67]

ABSTRACT

RNA transcripts are subject to post-transcriptional gene regulation involving hundreds of RNA-binding proteins (RBPs) and microRNA-containing ribonucleoprotein complexes (miRNPs) expressed in a cell-type dependent fashion. We developed a cell-based crosslinking approach to determine at high resolution and transcriptome-wide the binding sites of cellular RBPs and miRNPs. The crosslinked sites are revealed by thymidine to cytidine transitions in the cDNAs prepared from immunopurified RNPs of 4-thiouridine-treated cells. We determined the binding sites and regulatory consequences for several intensely studied RBPs and miRNPs, including PUM2, QKI, IGF2BP1-3, AGO/EIF2C1-4 and TNRC6A-C. Our study revealed that these factors bind thousands of sites containing defined sequence motifs and have distinct preferences for exonic versus intronic or coding versus untranslated transcript regions. The precise mapping of binding sites across the transcriptome will be critical to the interpretation of the rapidly emerging data on genetic variation between individuals and how these variations contribute to complex genetic diseases.

2.1 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 [10]. But Ago is just one out of hundreds of RNA-binding proteins (RBPs) and ribonucleoprotein complexes (RNPs) [132] that modulate the maturation, stability, transport, editing and translation of RNA transcripts in vertebrates [128, 136, 164]. 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 [84].

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) [172] 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 [29] or for Pumilio [53].

More direct RBP target site information is obtained by combining in vivo UV crosslinking [57, 183] with immunoprecipitation [33, 131] followed by the isolation of crosslinked RNA segments and cDNA sequencing (CLIP) [176]. CLIP was used to identify targets of the splicing regulators NOVA1 [121], FOX2 [198] and SFRS1 [152] as well as U3 snoRNA and pre-rRNA [56], pri-miRNA targets for HNRNPA1 [65], EIF2C2/AGO2 protein binding sites [24] and ALG-1 target sites in C. elegans [201]. 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.

2.2 RESULTS

2.2.1 *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 [93, 133]. Several of these photoactivatable nucleosides are readily taken up by cells without apparent toxicity and have been used for in vivo crosslinking [39]. We applied a subset of these nucleoside analogs (Figure 1A) 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

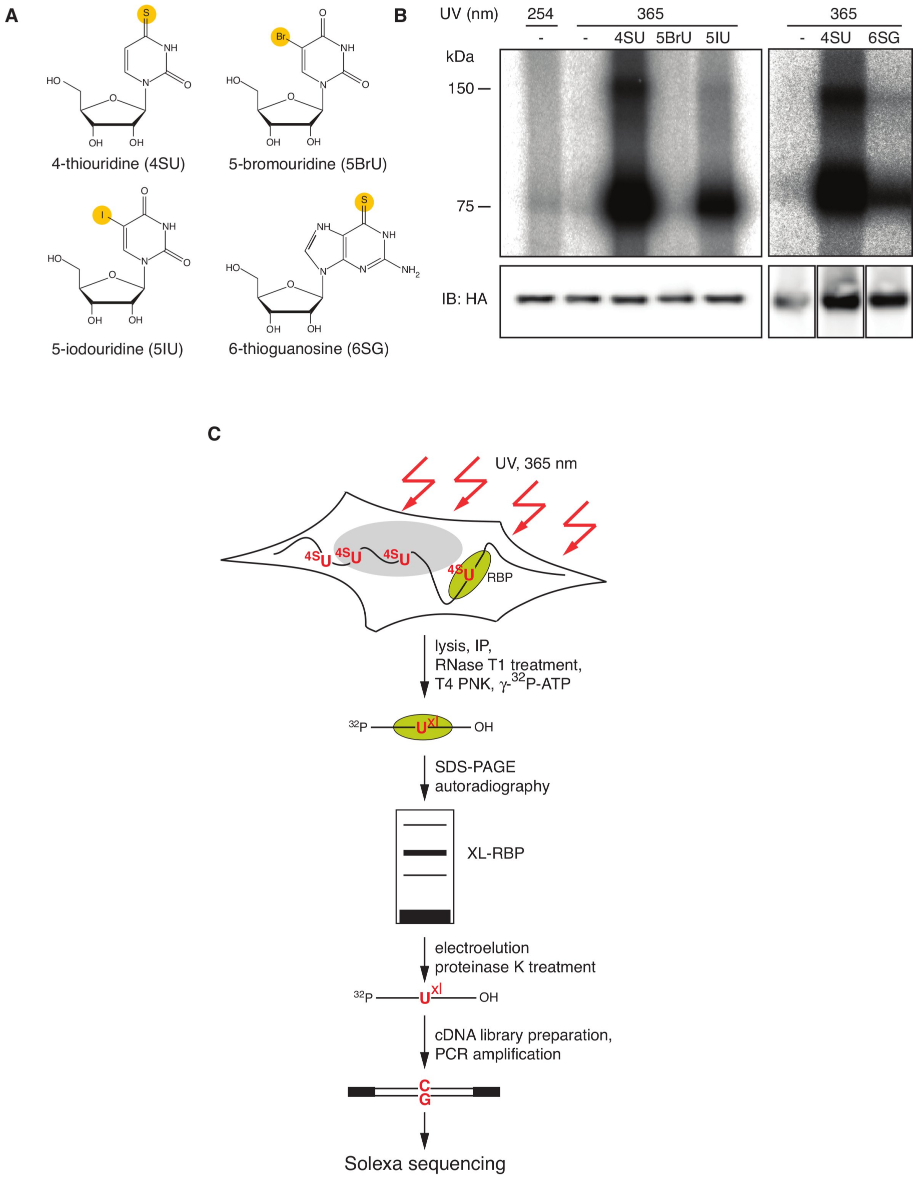


Figure 1: PAR-CLIP methodology (A) Structure of photoactivatable nucleosides (B) Phosphorimages of SDS-gels that resolved 5′-32P-labeled RNA-FLAG/HA-IGF2BP1 immunoprecipitates (IPs) prepared from lysates from cells that were cultured in media in the absence or presence of 100 µM photoactivatable nucleoside 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.

photoactivatable nucleosides improved RNA recovery 100- to 1000-fold, using the same amount of radiation energy (Figure 1B). We refer to our method as PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) (Figure 1C).

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 labelling 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 2A) known for its highly sequence-specific RNA binding [185].

2.2.2 *Identification of PUM2 mRNA targets and its RRE*

PUM2 protein crosslinked well to 4SU-labeled cellular transcripts (Figure 2B). The crosslinked segments were converted into a cDNA library and Solexa sequenced [66]. 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 2C, 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 20) in agreement with previous studies [190]. All sequence clusters with mapping and annotation information are available online[1].

PhyloGibbs analysis [160] of the top 100 most abundantly sequenced clusters (Table S3), as expected, yielded the PUM2 RRE, UGUA-NAUA [48] (Figure 2D). Unexpectedly, over 70% of all sequence reads that gave rise to clusters showed a T to C mutation compared to the genome (Figure 20). Ranking of sequence read clusters according to the frequency of T to C mutation further enriched for the PUM2 RRE (Figure 20) 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 T corresponding to U7 of the RRE mutated at higher frequency compared to the Ts corresponding to U1 and U3 (Figure 2E). 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.

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

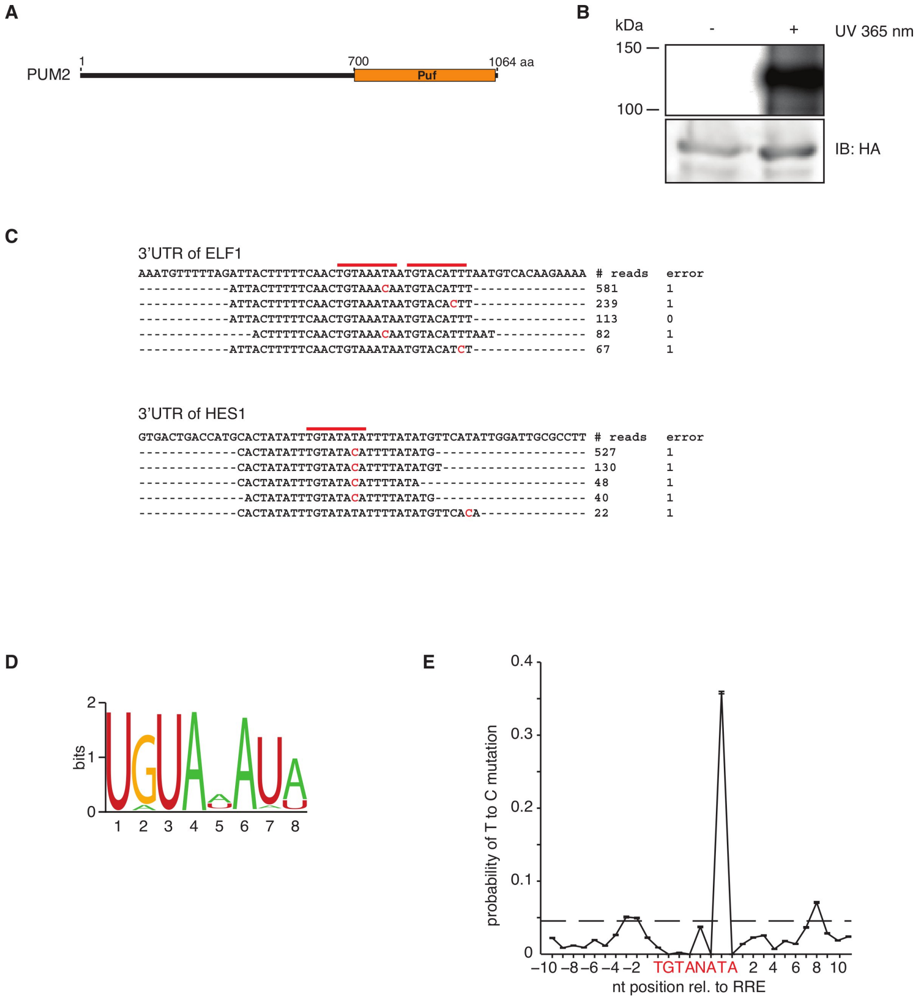


Figure 2: RNA recognition by PUM2 protein (A) Domain structure of PUM2 protein. (B) Phosphorimage of SDS-gel of radiolabeled FLAG/HA-PUM2-RNA complexes from non-irradiated or UV-irradiated 4SU-labeled cells. The lower panel shows an anti-HA immunoblot. (C) Alignments of PAR-CLIP cDNA sequence reads to 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 20.

2.2.3 *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 3A,B). The RRE ACUAAY was determined by SELEX [47], 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 [22].

PhyloGibbs analysis of the 100 most abundantly sequenced clusters (Table S3) yielded the RRE AYUAAY (Figures 3C,D), similar to a motif identified by SELEX [47]. 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 21).

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 3E). 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.

2.2.4 *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 3F,G). The crosslinking efficiency varied 50-fold and mirrored the results of the mutational analysis (Figure 3G). 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 indicated positions (Figure 3F). The fraction of sequence reads with T to C changes obtained from non-irradiated 4SU-containing oligori-

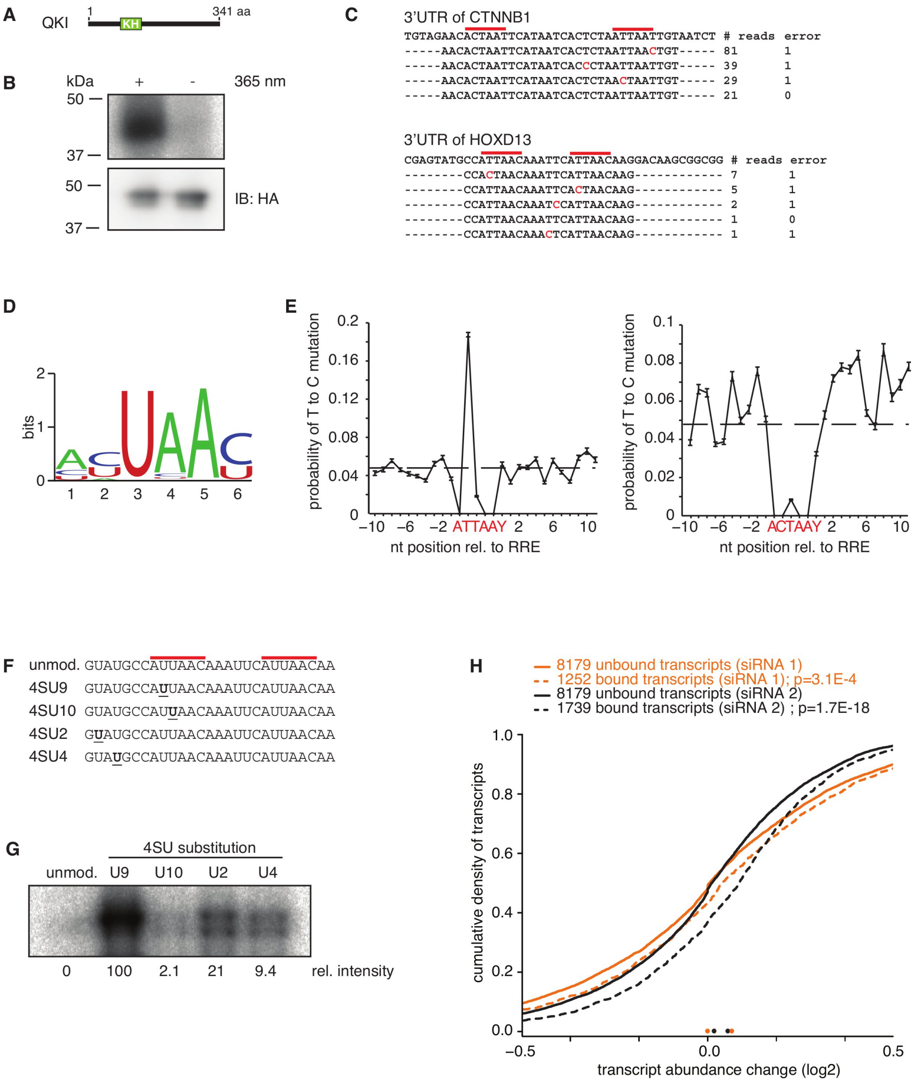


Figure 3: RNA recognition by QKI protein (A) Domain structure of QKI protein (B) Phosphorimage of SDS-gel resolving radiolabeled RNA crosslinked to FLAG/HA-QKI IPs from non-irradiated or UV-irradiated 4SU-labeled cells. The lower panel shows the anti-HA immunoblot. (C) Alignments of PAR-CLIP cDNA sequence 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 21.

bonucleotides 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 20). 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 3H), consistent with previous reports of QKI being a repressor [22].

2.2.5 *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 4A,B), a family of highly conserved proteins that play a role in cell polarity and cell proliferation [199]. These proteins are predominantly expressed in the embryo and regulate mRNA stability, transport and translation. They are re-expressed in various cancers [17, 30] and IGF2BP2 has been associated with type-2 diabetes [153]. The IGF2BPs are highly similar and contain six canonical RNA-binding domains, two RNA recognition motifs (RRMs) and four KH domains (Figure 4A). 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 [199].

The three IGF2BPs recognized a highly similar set of target transcripts (Table S1), suggesting similar and redundant functions. Phy-

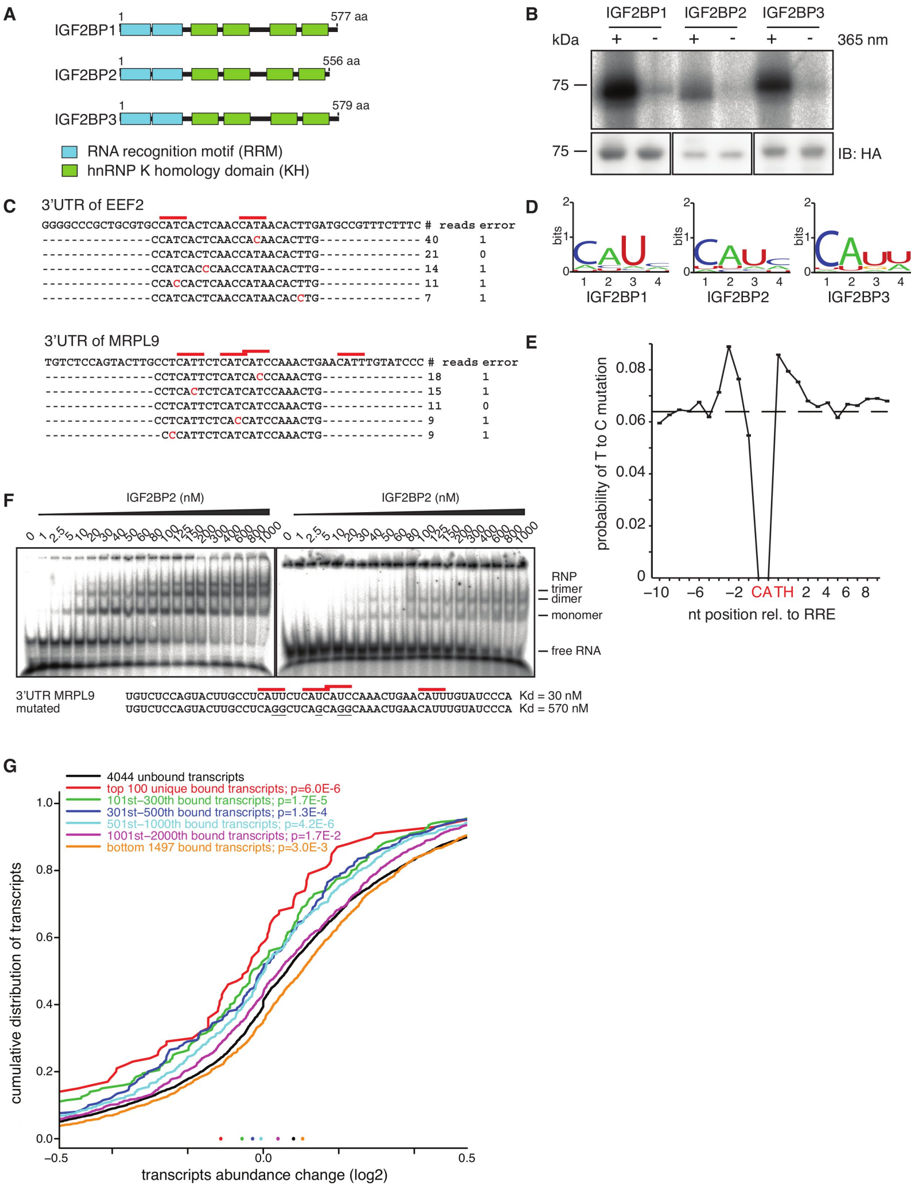


Figure 4: RNA recognition by the IGF2BP protein family (A) Domain structure of IGF2BP1-3 proteins. (B) Phosphorimage of an SDS-gel resolving radiolabeled RNA crosslinked to FLAG/HA-IGF2BP1-3 IPs. The lower panel shows anti-HA immunoblots. (C) Alignments of IGF2BP1 PAR-CLIP cDNA sequence reads to the 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 22 and 23.

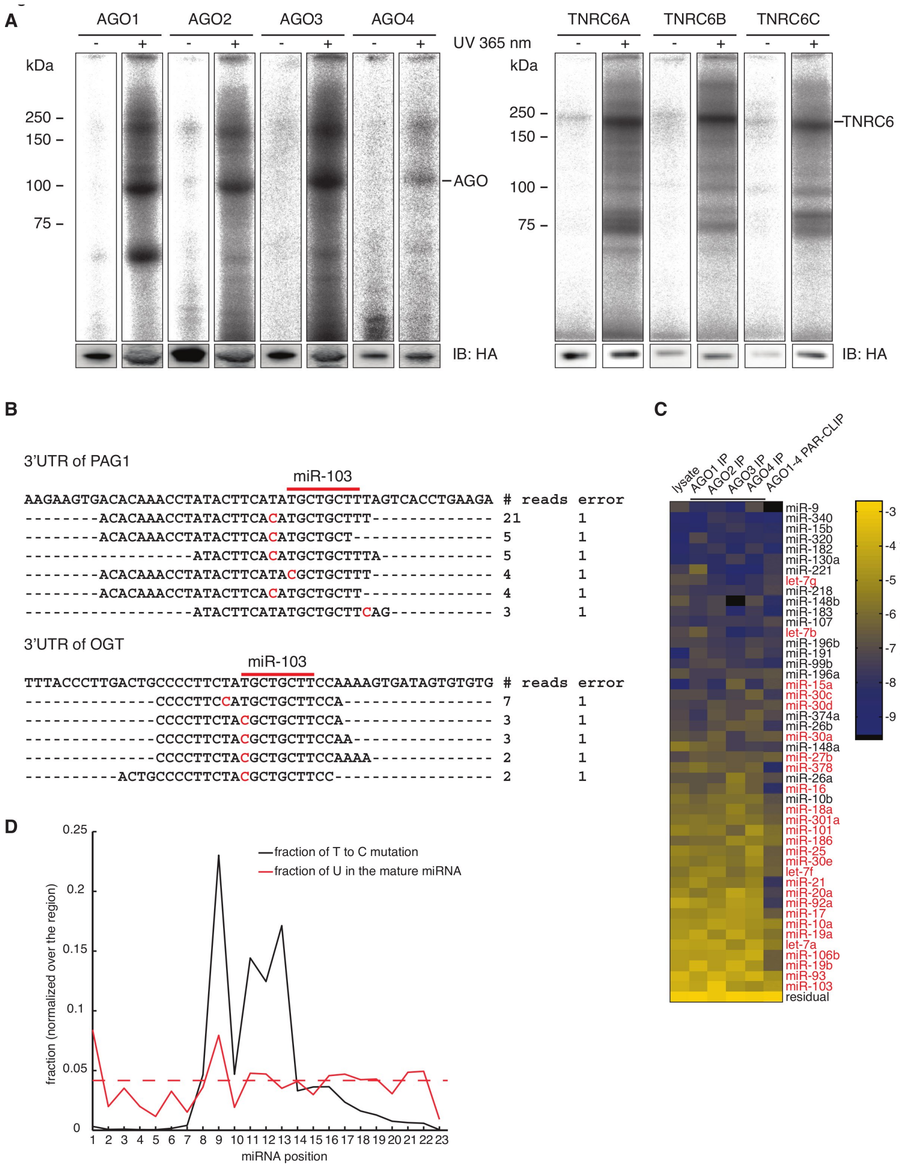
loGibbs analysis of the clusters derived from mRNAs (Figure 4C and Table S3) yielded the sequence CAUH (H=A, U, or C) as the only consensus recognition element (Figure 4D), contained in more than 75% of the top 1000 clusters for IGF2BP1, 2 or 3 (Figure 22). In total, we identified over 100,000 sequence 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 22). 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 4E). 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 22). In vitro binding assays showed that nucleotide changes of the CAUH motif decreased, but did not abolish the binding affinity (Figure 4F and Table S1).

To test the influence of IGF2BPs on the stability of their interacting mRNAs, as reported previously for some targets [199], 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 4G), 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 [121, 176], 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 23).

2.2.6 *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") [10]. Target sites were thought to be predominantly located in the 3'UTRs of mRNAs, and computational miRNA target prediction methods frequently resort to identification



3'UTR of PAG1

miR-103

```
AAGAAGTGACACAAACCTATACTTCATATGCTGCTTTAGTCACCTGAAGA # reads error
--------ACACAAACCTATACTTCACATGCTGCTTT------------- 21      1
--------ACACAAACCTATACTTCACATGCTGCT--------------- 5       1
------------------ATACTTCACATGCTGCTTTA------------ 5       1
--------ACACAAACCTATACTTCATACGCTGCTTT------------- 4       1
--------ACACAAACCTATACTTCACATGCTGCTT-------------- 4       1
------------------ATACTTCATATGCTGCTTCAG----------- 3       1
```

3'UTR of OGT

miR-103

```
TTTACCCTTGACTGCCCCTTCTATGCTGCTTCCAAAAGTGATAGTGTGTG # reads error
--------------CCCCTTCCATGCTGCTTCCA---------------- 7       1
--------------CCCCTTCTACGCTGCTTCCA---------------- 3       1
--------------CCCCTTCTACGCTGCTTCCAA--------------- 3       1
--------------CCCCTTCTACGCTGCTTCCAAAA------------- 2       1
----------ACTGCCCCTTCTACGCTGCTTCC----------------- 2       1
```

Figure 5: 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 24.

of evolutionarily conserved sites that are located in 3'UTRs and are complementary to miRNA seed regions [10, 146]. We isolated mRNA fragments bound by miRNPs from HEK293 cell lines stably expressing FLAG/HA-tagged AGO or TNRC6 family proteins [106]. 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 5A).

From clusters (Figure 5B) 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.

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 [106]. 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 [106]. 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 24).

2.2.7 *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 [66], 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 5C and Table S5) supporting our observation that AGO1-4 bind the same targets [106]. 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 5C and Figure 24A).

Importantly, in the AGO PAR-CLIP library, the majority of miRNA sequence reads derived from prototypical miRNAs [104] displayed T to C conversion near or above 50%. The T to C conversion was predominantly concentrated within positions 8 to 13 (Figure 5D), residing in the unpaired regions of the AGO protein ternary complex [186]. 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 24).

2.2.8 *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 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 6A). 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 [186]. 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 [186, 187]. 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 [61, 73].

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 6B). 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.

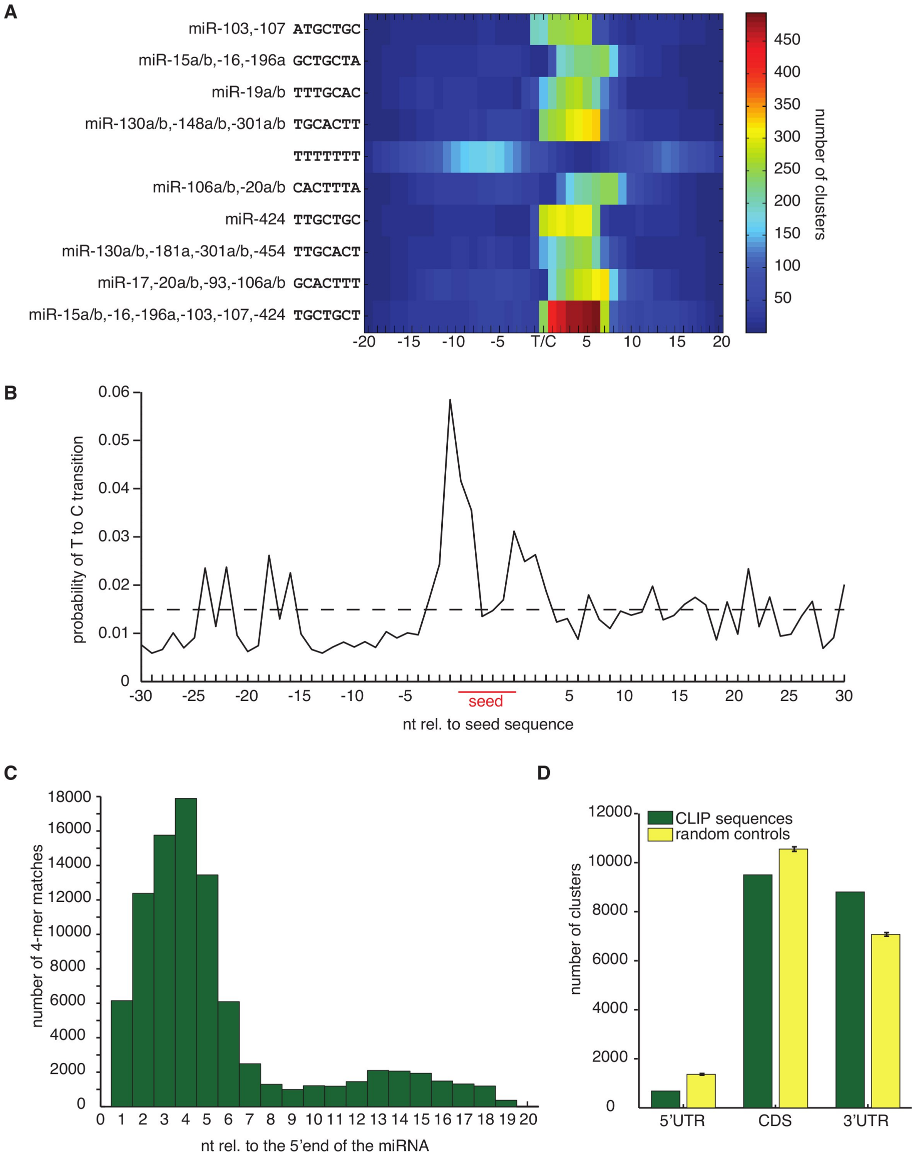


Figure 6: AGO PAR-CLIP identifies miRNA seed-complementary sequences in HEK293 cells. (A) Representation of the 10 most significantly enriched 7-mer sequences within PAR-CLIP CCRs. T/C indicates the predominant T to C transition within clusters of sequence reads. (B) T to C positional mutation frequency 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 25.

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 24), 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 [46, 116] (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 25B). 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 25C), which was expected from previous bioinformatic analyses of functional miRNA-binding sites.

2.2.9 *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 [186]. 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 [61], we also detected a weak signal for contiguous 4-nt long matches to positions 13-15 of the miRNA (Figure 6C).

2.2.10 *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 6D).

Evidence of widespread binding of miRNAs to the CDS was reported before [35, 116]. However, miRNAs are believed to predominantly act on 3′UTRs [10], with relatively few reports providing experimental evidence for miRNA-binding to individual 5′UTRs or CDS [35, 42, 126, 142, 171]. To obtain evidence that AGO CCRs indeed contain functional miRNA-binding sites, we blocked 25 of the most abundant miRNAs in HEK293 cells (Figure 5C) by transfection of a cocktail of 2′-O-methyl-modified antisense oligoribonucleotides and monitored the changes in mRNA stability by microarrays (Figure 7A). Consistent with previous studies of individual miRNAs [61], 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 7B). 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 7C), and that multiple binding sites may also reside within a single 41-nt CCR (Figure 25). Both of these findings are in agreement with previous observations [61]. 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 7D), independent of the extent of seed pairing. To gain more insights into miRNA binding in the CDS, we examined the codon adaptation index (CAI) [158] 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- as well as downstream of the crosslinked seed matches (Figure 7E), 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

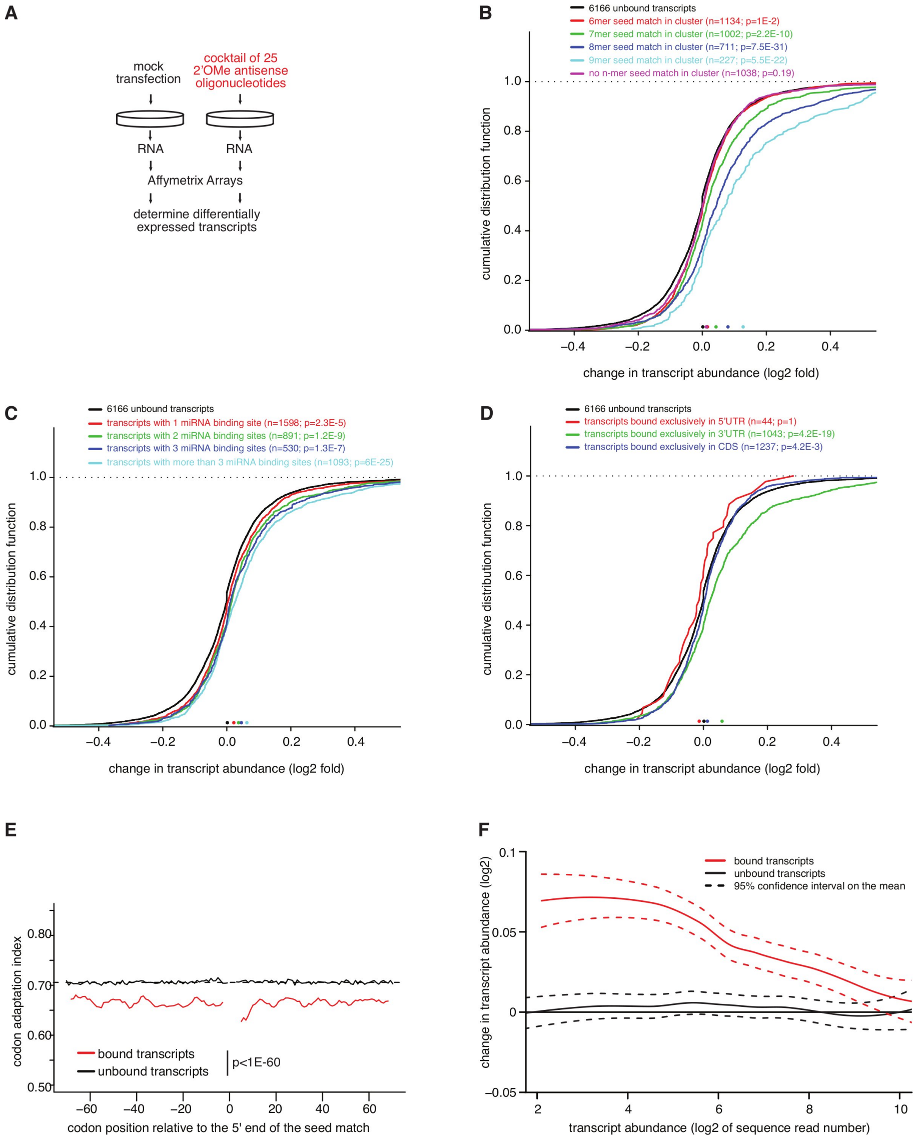


Figure 7: Relationship between various features of miRNA/target RNA interactions and mRNA stability (A) FLAG/HA-AGO2-tagged HEK293 cells were transfected with a cocktail of 25 2′-O-methyl modified antisense oligoribonucleotides, inhibiting miRNAs marked in red in Figure 5C, or mock transfected, followed 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 26.

miRNA-target-RNA interactions as ribosomes are stalled [64]. 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 7F). Highly expressed mRNAs appear to avoid miRNA regulation [165], 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 [9, 122, 156]. We transfected miRNA duplexes corresponding to the deeply conserved miR-7 and miR-124 into FLAG/HA-AGO2 cells, performed PAR-CLIP (Figure 26), 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 26).

2.2.11 *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 [46], the mRNA stability scores by TargetScan context score [61], 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 26), 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 26). 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 26).

Finally, we wanted to examine how miRNA targets defined by PAR-CLIP compared in regulation of target mRNA stability to those predicted by ElMMo [46], TargetScan context score [61], TargetScan Pct [44] and PicTar [103]. 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 26).

2.3 DISCUSSION

Maturation, localization, decay and translational regulation of mRNAs involve formation of complexes of RBPs and RNPs with their RNA targets [128, 136]. 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 [125]. Furthermore hundreds of miRNAs function together with AGO and TNRC6 proteins to destabilize target mRNAs and/or repress their translation [10]. 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.

2.3.1 *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 [24, 121, 176, 201], except for recent work by Granneman et al. [56] 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 22).

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

2.3.2 *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.

2.3.3 *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 [64].

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 [61], 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 [182] and those recently observed in biochemical and structural studies of T. thermophilus Ago protein [186, 187].

2.3.4 *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.

2.4 METHODS

2.4.1 *PAR-CLIP*

Human embryonic kidney (HEK) 293 cells stably expressing FLAG/HA-tagged IGF2BP1-3, QKI, PUM2, AGO1-4, and TNRC6A-C [106] 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 [66] 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.

2.4.2 *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 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 in the Supplementary Material.

2.4.3 *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.

2.5 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.

2.6 CONFLICT OF INTEREST STATEMENT.

Thomas Tuschl is a co-founder and scientific advisor to *Alnylam Pharmaceuticals* and an advisor to *Regulus Therapeutics*.

3

A DATABASE AND ANALYSIS ENVIRONMENT FOR EXPERIMENTALLY-DETERMINED BINDING SITES OF RNA-BINDING PROTEINS

The work presented in this chapter was originally published in Nucleic Acids Research [90]

ABSTRACT

The stability, localization and translation rate of mRNAs are regulated by a multitude of RNA-binding proteins (RBPs) that find their targets directly or with the help of guide RNAs. Among the experimental methods for mapping RBP binding sites, crosslinking and immunoprecipitation (CLIP) coupled with deep sequencing provides transcriptome-wide coverage as well as high resolution. However, partly due to their vast volume, the data that were so far generated in CLIP experiments have not been put in a form that enables fast and interactive exploration of binding sites. To address this need, we have developed the CLIPZ database and analysis environment. Binding site data for RBPs such as Argonaute 1-4, Insulin-like growth factor II mRNA-binding protein 1-3, TNRC6 proteins A-C, Pumilio 2, Quaking, and Polypyrimidine tract binding protein can be visualized at the level of the genome and of individual transcripts. Individual users can upload their own sequence data sets while being able to limit the access to these data to specific users, and analyses of the public and private data sets can be performed interactively. CLIPZ, available at http://www.clipz.unibas.ch aims to provide an open-access repository of information for post-transcriptional regulatory elements.

3.1 INTRODUCTION

Almost all cellular RNAs interact with RNA-binding proteins (RBPs) to form ribonucleoprotein complexes (RNPs). The overall composition and precise architecture of these RNPs undergo dynamic remodeling in response to signals and cellular state. Initial annotation [79] indicated that the human genome contains approximately 300 genes that encode proteins with an RNA-recognition motif (RRM). This is only one of the over 40 distinct protein domains known to contact RNA. RBP-RNA interactions are highly context dependent and many RBPs carry out different functions in different cellular compartments. For instance, the T-cell intracellular antigen 1 (TIA-1) functions as a splicing factor in the nucleus; it binds to an intronic splice enhancer in the Fas pre-mRNA leading to the inclusion of the proximal exon [179]. In the cytoplasm, TIA-1 regulates the stability of mature mRNAs: its binding to AU-rich elements located in the 3′ untranslated regions (3′UTRs) of mRNAs (such as that of transforming growth factor beta, TGFβ) attracts the mRNA degradation machinery. The same AU-rich element in the TGFβ 3′UTR when bound by the HuR RBP leads to mRNA stabilization [179]. Thus, precise knowledge of spatio-temporal associations between RBPs and mRNAs under various conditions is key to understanding how the level, translation rate and cellular localization of those mRNAs are regulated during the life time of a cell.

With some exceptions, such as the knowledge-based potential function designed by Zheng et al. [200] to predict the specificity and relative binding energy of RNA-binding proteins, computational models describing the binding specificity of RBPs (by contrast, for instance, with transcription factors) are lacking [7]. Recently however, experimental methods for high-throughput and high-resolution identification of RBP binding sites have been developed. They rely on crosslinking and immunoprecipitation (CLIP) of RBPs of interest [176] followed by deep sequencing [25, 67, 96]. In a particular variant of CLIP, termed PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation), the incorporation of photo-reactive nucleotides in mRNAs prior to crosslinking induces a specific mutational signature in the sequenced reads relative to the reference genome, thereby enabling the separation of crosslinked binding sites from other RNA fragments that are captured non-specifically during the experiment [67]. Many questions concerning the function, specificity and modulation of activity of RBPs can be addressed through analyses of corresponding PAR-CLIP data sets. For example, the sites with the highest number of crosslinking events (indicated by T-to-C mutations in the sequenced reads) can be analyzed to uncover the sequence specificity of the RBP and to identify cellular pathways that are targeted by the RBP in a concerted manner. Moreover, with PAR-CLIP data available for multiple RBPs, one can begin to identify regions of crosstalks between multiple RBPs on individual mRNAs.

Here we describe a database of binding sites that we constructed based on CLIP data for various proteins that are known to regulate mRNA splicing (Polypyrimidine tract binding protein), stability and/or translation rate (Quaking, Pumilio2, Argonautes 1-4, TNRC6 A-C, Insulin-like growth factor II mRNA-binding proteins 1-3). The data is presented through a web interface that supports not only visualizations but also further analyses of RBP binding sites. The platform also allows registered users to submit for functional annotation short reads resulting from CLIP, small RNA sequencing and mRNA sequencing experiments. Once uploaded, these data can be explored through various interactive analysis tools that we developed. Due to its user- and dataset-management system, the platform can support collaborative projects involving private and public data and multiple

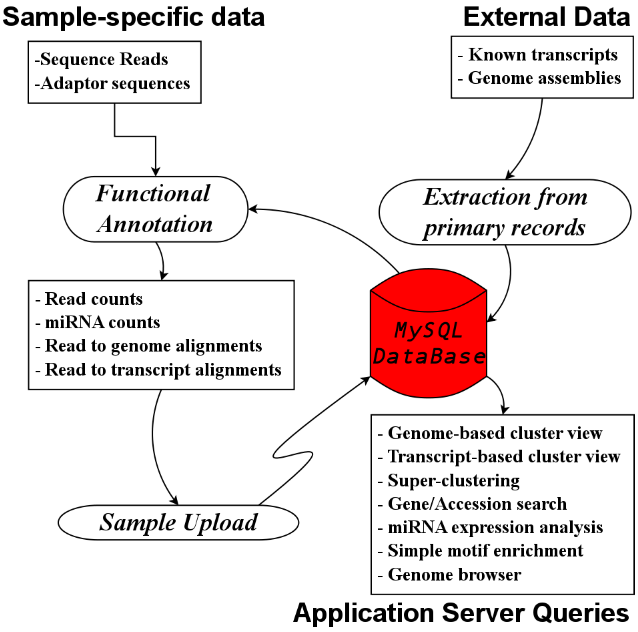


Figure 8: CLIPZ Data flow. Procedures are further described in the Methods section.

users. This resource is of great value to researchers that study the mechanisms regulating mRNA stability and translation.

3.2 MATERIALS AND METHODS

3.2.1 *Sequence annotation*

The computational pipeline underlying the construction of the CLIPZ database takes as input fasta-formatted files of sequences that were obtained from CLIP samples through deep sequencing. These sequences are submitted to an initial annotation process that attempts to identify the origin (within the genome and within known transcripts) of individual sequence reads. The annotation procedure is described in detail elsewhere [14]. Briefly, it consists of adaptor removal, mapping of sequenced reads to the genome and to known transcripts, and functional annotation of each read based on its best mappings. A sketch of the data flow is shown in Figure 8.

3.2.1.1 *Adaptor removal.*

During sample preparation, adaptors are ligated at both 5′ and 3′ ends of CLIP sequence fragments. Because most of the CLIP data that is currently available has been generated using the Solexa sequencing technology [13], our procedure for adaptor removal is specific to this technology (though other adaptor configurations can easily be taken into account). The 5′ adaptor serves as a sequencing primer, and we expect that only the 3′ adaptor (or part of it) is sequenced. We use an in-house ends-free local alignment algorithm [77] (parameters: 2 for match, −3 for mismatch, −5 for gap opening, −2 for gap extension) to align the 3′ adaptor to the reads. The part of the sequence read that aligns to the 5′ end of the 3′ adaptor is removed, and if the remainder of the read is longer than 15 nucleotides, it is retained for further analysis. Distinct sequences are deposited in the database together with their copy number in the sample under study.

3.2.1.2 *Sequence mapping.*

All distinct sequences are mapped to the genome assembly. Currently, the database contains CLIP samples obtained from human cells, for which we used the *hg18* version of the human genome assembly from the University of California at Santa Cruz [1], but analyses of mouse data sets are also supported. Because not all transcripts that have been sequenced and are present in sequence databases can be mapped to the genome assembly and because various contaminants can be found in CLIP samples, we also map the reads to a database of sequences with known function (ribosomal, transfer, small cytoplasmic, small nuclear and small nucleolar RNAs, PIWI proteins-associated RNAs, miRNAs, messenger RNAs, miscellaneous non-coding RNAs obtained from sequencing projects, bacterial and fungal ribosomal RNAs, genomes of common bacteria, vector, adaptor, and size marker sequences). The sources of these sequences are as follows.

- **Protein-coding as well as non-coding sequences** (mRNA, tRNA, scRNA, snoRNA, rRNA, snRNA, piRNA, and miscRNA) were extracted based on the *molecule type* and *feature key* fields of the Genbank records (Genbank release of January 19, 2010).

- **Bacterial and fungal sequences** (ribosomal RNA and complete genome sequences of bacterial and fungal species available in the Nucleotide Database of Genbank) were included in order to detect contaminations.

- **Vector sequences** [2] were included for the same purpose.

- **miRNAs precursor sequences** [3] were included to be able to analyze data specific to proteins involved in the miRNA-dependent silencing pathway.

- **Miscellaneous non-coding RNAs** [4] were included to enable detection of interactions involving poorly characterized non-coding RNAs, that have not been categorized yet in Genbank databases.

[1] http://genome.cse.ucsc.edu

[2] ftp://ftp.ncbi.nih.gov/pub/UniVec/UniVec

[3] ftp://mirbase.org/pub/mirbase/CURRENT

[4] http://www.noncode.org

- **Human tRNAs** [5]
- **Mouse tRNAs** [6] were included because these resources provide extensive annotations of tRNAs.
- **Repeat Masker annotations** of the genome provided by the University of California at Santa Cruz [7] were used to detect repeat elements.

To align sequence reads to target sequences, we use the *Oligomap* algorithm [14] that exhaustively reports the mappings with 0 or 1 error (mismatch, insertion or deletion). The *Oligomap* software can be downloaded from http://www.mirz.unibas.ch/software.php. In principle, we take into account all the possible loci for a given sequence read and we assume that the read originated from any of these loci with equal probability. Based on the *GMAP* [193] mappings of mRNAs to genome, we determine whether a genome-mapped read falls inside an intronic or exonic region. Based on the coding region annotation of transcripts in Genbank, we determine whether the exonic sequence reads originate from the 5'UTR, CDS or 3'UTR region of individual transcripts. Sequence reads that map to regions with alternative splicing are assigned fractional numbers that denote the proportion of transcripts in which the region appeared in a particular section of the transcript.

3.2.1.3 *Sequence annotation.*

Whenever an extracted sequence read maps to one or more known sequence(s) of the same functional category, that functional category is readily transferred to the sequence read. There are however, sequence reads that map equally well to known sequences of different functional categories (e.g. tRNA, rRNA, mRNA and repeat). In these cases we assign a functional annotation with the following priority scheme rRNA > tRNA > snRNA > snoRNA > scRNA > miRNA > piRNA > repeat > miscRNA > mRNA (reflecting roughly the abundance of various types of sequences in the cell).

3.2.2 *Generation of clusters of sequence reads*

Initial analysis of PAR-CLIP data indicated that the sequence reads obtained in individual experiments generally form well-delimited, relatively short (20-40 nucleotides) clusters. When the binding specificity of the protein was already known, the clusters obtained from PAR-CLIP data typically *contained* the sequence motif known to represent the binding site of the protein [67]. We therefore use a cluster as the

[5] http://lowelab.ucsc.edu/GtRNAdb/Hsapi/Hg17-tRNAs.fa

[6] http://lowelab.ucsc.edu/GtRNAdb/Mmusc/Mm6-tRNAs.fa

[7] http://genome.cse.ucsc.edu

central unit for data analysis and visualization. Two sequences are placed in the same cluster if they overlap by at least one nucleotide in their genomic or transcript location). We note that in data sets obtained with other CLIP protocols, the correspondence between clusters that are generated this way and individual RBP binding sites may not be as clear as it is in PAR-CLIP. As more data generated with different variants of CLIP becomes available, the definition of the visualization unit (ideally the RBP binding site) may need to be revised accordingly. Furthermore, in PAR-CLIP experiments T-to-C mutations are indicative of crosslinked positions and our analysis has shown that clusters with the largest number of T-to-C mutations are most enriched in functional binding sites for the studied RBP. The number of T-to-C mutations as well as other statistics are therefore computed for each cluster and made available in the interface. The user can sort the clusters based on these computed features in order to extract the targets that are most frequently bound by the RBP of interest.

3.2.3 *Data storage*

We use a *MySQL 5* database management system to store the results of the functional annotation process and to support downstream analyses. The database contains the following types of tables:

- **User-Management-related:** tables that store information about the user (name, the group in which he/she is a member, the host laboratory, etc.)

- **Known-Sequence-related:** tables that contain information about transcripts of known function that we obtained from external sources and used for short read annotation (the sequences themselves, genomic loci, NCBI Entrez Gene information when available, etc.).

- **Sample-Data-related:** tables that contain information about the sequences from a submitted sample (e.g. extracted sequence reads, genomic loci, mapping coordinates within transcripts of known function, etc.).

- **Sequence-Read-Cluster-related:** tables that contain information about clusters of overlapping sequences typically representing individual binding sites. The cluster information is used in various visualizations.

In order to maximize the efficiency of processing subsequent queries, database tables are generated for each individual sample (for the detailed description of the database schema see the "Help" pages provided on the web site).

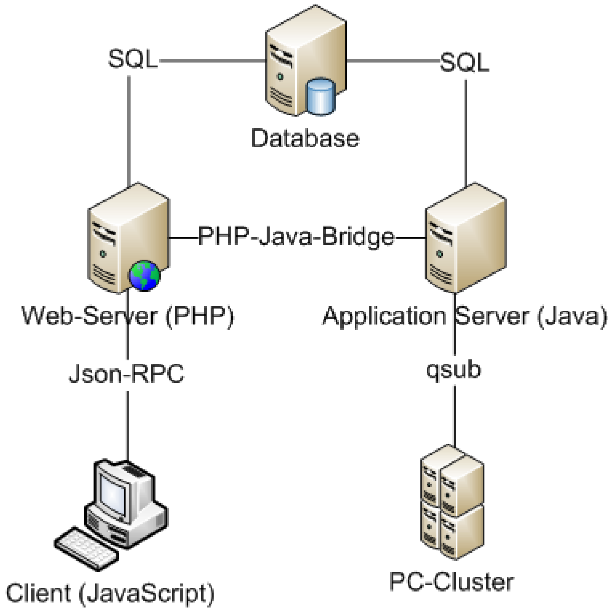


Figure 9: Architecture of the software underlying the CLIPZ analysis environment.

3.2.4 *Analysis environment*

The software supporting the web-based queries has the following components (see Figure 9).

3.2.4.1 ***Web Server***

The web server is responsible for the validation of the user inputs and for rendering the results of various computations. It uses *PHP 5* and a *Model View Controller (MVC)-Framework* that we developed. It communicates with the application server using a freely available PHP-Java bridge [8].

3.2.4.2 ***Application Server***

The application server, implemented in *Java 1.6*, provides functions that can be accessed by the web server, such as applying the functional annotation pipeline to an uploaded sample. It is also responsible for process control, logging the job outputs and reporting the errors whenever jobs fail. Due to the large volume of typical CLIP data sets, we employ a *PC-Cluster* for parallel processing. The job distribution to the cluster and the handling of conflicts that may result from multiple parallel-running jobs requesting the same data/resource at the same time are also handled by the application server.

[8] http://php-java-bridge.sourceforge.net/pjb/

3.2.4.3 *Client*

User-interactivity is provided by various JavaScript libraries such as:

- Dojo [9]
- YUI [10]
- JQuery [11]

We established a Generic Genome Browser [166] server to generate transcript- or genome-based views of the location of binding sites for one or more proteins in the data set. Communication between the JavaScript on the client side and the web server is being established with a Remote Procedure Call (RPC) System that we developed based on the *JsonRPC 2.0* protocol. This is described at http://groups.google.com/group/json-rpc/web/json-rpc-1-2-proposal.

3.3 EXAMPLES OF INTERACTIVE ANALYSES

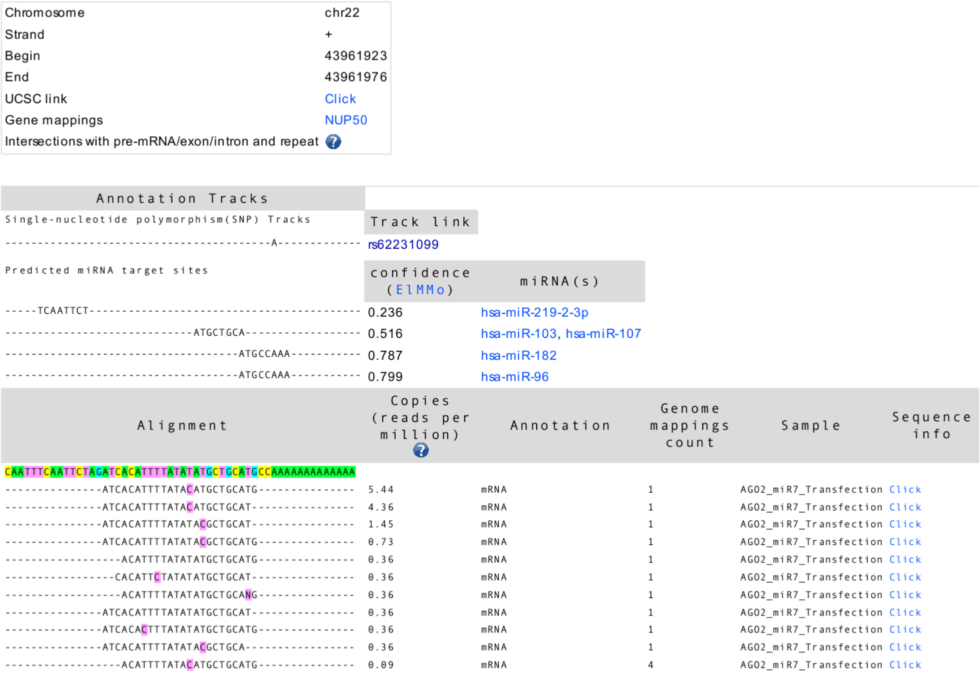
3.3.1 *Visualization of clusters of genome- or transcript-based clusters of reads*

For each sample in the database, the user can browse the clusters of overlapping sequence reads which in the PAR-CLIP samples typically correspond to individual RBP binding sites. The clusters can be sorted by various criteria including the number of T-to-C mutations in all reads of a cluster, which in the PAR-CLIP experiments is indicative of the affinity of the protein for the RNA. To distinguish crosslink-induced mutations from single nucleotide polymorphisms (SNPs) we incorporated a track that shows the known SNPs, and for identifying the miRNAs that guide the Argonaute to the target RNA, we incorporated a track of predicted miRNA binding sites [46] (see Figure 10).

3.3.2 *Transcript and genome browsers*

The association of an RBP with a specific site and the downstream effects of this interaction frequently depend on other regulatory elements that are present in close vicinity and recruit other regulatory factors. Through the transcript and genome browsers, one can visualize the position of binding sites within transcripts, as well as the spatial relationship between binding sites determined in different experiments, as shown in Figure 11.

[9] http://www.dojotoolkit.org
[10] http://developer.yahoo.com/yui
[11] http://www.jquery.com



Chromosome	chr22
Strand	+
Begin	43961923
End	43961976
UCSC link	Click
Gene mappings	NUP50
Intersections with pre-mRNA/exon/intron and repeat	?

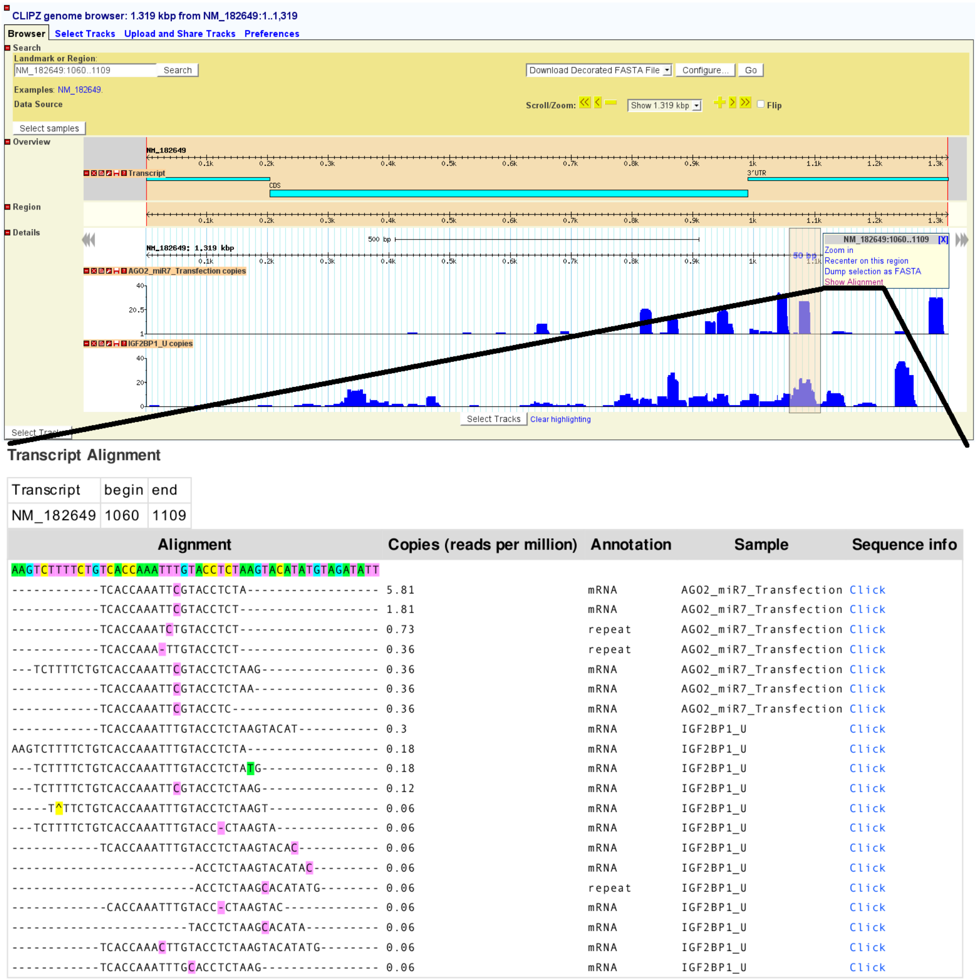
Annotation Tracks		
Single-nucleotide polymorphism(SNP) Tracks	Track link	
---A------------	rs62231099	
Predicted miRNA target sites	confidence (ElMMo)	miRNA(s)
-----TCAATTCT---	0.236	hsa-miR-219-2-3p
-----------------------------ATGCTGCA-----------------	0.516	hsa-miR-103, hsa-miR-107
------------------------------------ATGCCAAA----------	0.787	hsa-miR-182
------------------------------------ATGCCAAA----------	0.799	hsa-miR-96

Alignment	Copies (reads per million) ?	Annotation	Genome mappings count	Sample	Sequence info
CAATTTCAATTCTAGATCACATTTTATATATGCTGCATGCCAAAAAAAAAAAAA					
---------------ATCACATTTTATACATGCTGCATG---------------	5.44	mRNA	1	AGO2_miR7_Transfection	Click
---------------ATCACATTTTATACATGCTGCAT----------------	4.36	mRNA	1	AGO2_miR7_Transfection	Click
---------------ATCACATTTTATATACGCTGCAT----------------	1.45	mRNA	1	AGO2_miR7_Transfection	Click
---------------ATCACATTTTATATACGCTGCATG---------------	0.73	mRNA	1	AGO2_miR7_Transfection	Click
------------------ACATTTTATATATGCTGCATG---------------	0.36	mRNA	1	AGO2_miR7_Transfection	Click
-----------------CACATTCTATATATGCTGCAT----------------	0.36	mRNA	1	AGO2_miR7_Transfection	Click
------------------ACATTTTATATATGCTGCANG---------------	0.36	mRNA	1	AGO2_miR7_Transfection	Click
---------------ATCACATTTTATATATGCTGCAT----------------	0.36	mRNA	1	AGO2_miR7_Transfection	Click
---------------ATCACACTTTATATATGCTGCATG---------------	0.36	mRNA	1	AGO2_miR7_Transfection	Click
---------------ATCACATTTTATATACGCTGCA-----------------	0.36	mRNA	1	AGO2_miR7_Transfection	Click
------------------ACATTTTATACATGCTGCATG---------------	0.09	mRNA	4	AGO2_miR7_Transfection	Click

Figure 10: "Cluster View" of the AGO2/EIF2C PAR-CLIP reads mapping to the nucleoporin 50kDa (NUP50) gene, also showing single-nucleotide polymorphisms (SNPs) and predicted miRNA binding sites (with their corresponding probabilities given by the ElMMo model [46]) in the neighborhood of the RBP binding site.

3.3.2.1 *Genome/Known Sequence Super Clustering*

Many questions arising in the context of analyzing RBP binding sites can be phrased in terms of the spatial relationship between binding sites obtained in different experiments. For example, one would like to know whether experimental results for one protein are reproducible, in which case we expect that the sets of sites obtained in different experiments are largely identical. Alternatively, one may like to find out whether two proteins frequently compete for sites, in which case we would expect that the sites are occupied by one of the proteins in one condition and by the other protein in a different condition. The super-clustering tool enables the user to uncover such relationships. The visualizations that can be performed are very similar to those described for clusters of a single RBP but they operate on super-clusters that are built through single-linkage clustering of clusters obtained in different experiments and are either overlapping or at a specified maximum distance from each other. The user may define complex operations between sites obtained in different experiments using logical operators such as *(OR, AND, NOT)*.



CLIPZ genome browser: 1.319 kbp from NM_182649:1..1,319

Landmark or Region: NM_182649:1060..1109

Examples: NM_182649.

Scroll/Zoom: Show 1.319 kbp

☐ Flip

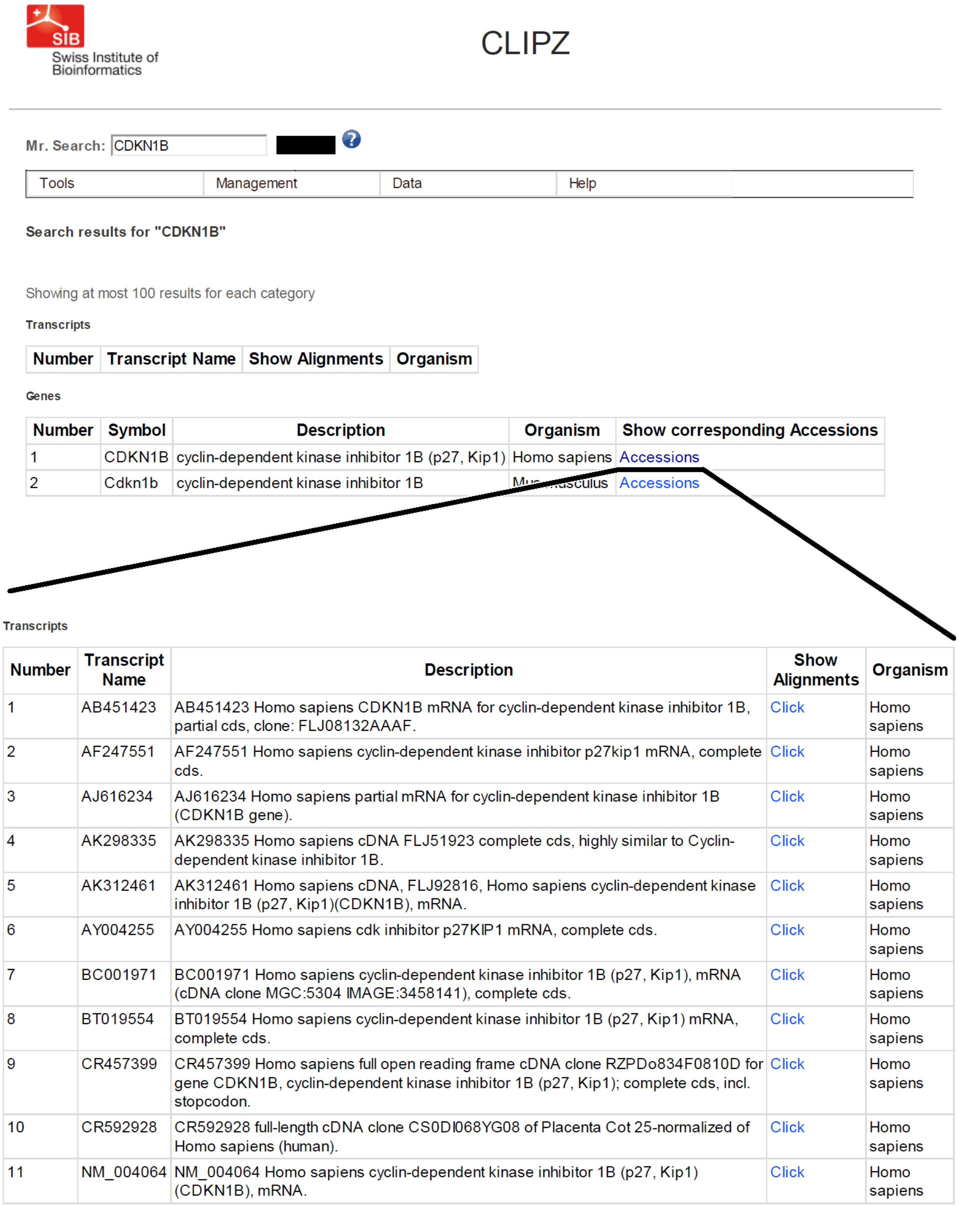
NM_182649:1060..1109 [X]

Transcript Alignment

Transcript	begin	end
NM_182649	1060	1109

Alignment	Copies (reads per million)	Annotation	Sample	Sequence info
AAGTCTTTTCTGTCACCAAATTTGTACCTCTAAGTACATATGTAGATATT				
------------TCACCAAATTCGTACCTCTA------------------	5.81	mRNA	AGO2_miR7_Transfection	Click
------------TCACCAAATTCGTACCTCT-------------------	1.81	mRNA	AGO2_miR7_Transfection	Click
------------TCACCAAATCTGTACCTCT-------------------	0.73	repeat	AGO2_miR7_Transfection	Click
------------TCACCAAA-TTGTACCTCT-------------------	0.36	repeat	AGO2_miR7_Transfection	Click
---TCTTTTCTGTCACCAAATTCGTACCTCTAAG----------------	0.36	mRNA	AGO2_miR7_Transfection	Click
------------TCACCAAATTCGTACCTCTAA-----------------	0.36	mRNA	AGO2_miR7_Transfection	Click
------------TCACCAAATTCGTACCTC--------------------	0.36	mRNA	AGO2_miR7_Transfection	Click
------------TCACCAAATTTGTACCTCTAAGTACAT-----------	0.3	mRNA	IGF2BP1_U	Click
AAGTCTTTTCTGTCACCAAATTTGTACCTCTA------------------	0.18	mRNA	IGF2BP1_U	Click
---TCTTTTCTGTCACCAAATTTGTACCTCTATG----------------	0.18	mRNA	IGF2BP1_U	Click
---TCTTTTCTGTCACCAAATTCGTACCTCTAAG----------------	0.12	mRNA	IGF2BP1_U	Click
-----T^TTCTGTCACCAAATTTGTACCTCTAAGT---------------	0.06	mRNA	IGF2BP1_U	Click
---TCTTTTCTGTCACCAAATTTGTACC-CTAAGTA--------------	0.06	mRNA	IGF2BP1_U	Click
------------TCACCAAATTTGTACCTCTAAGTACAC-----------	0.06	mRNA	IGF2BP1_U	Click
-------------------------ACCTCTAAGTACATAC---------	0.06	mRNA	IGF2BP1_U	Click
-------------------------ACCTCTAAGCACATATG--------	0.06	repeat	IGF2BP1_U	Click
-------------CACCAAATTTGTACC-CTAAGTAC-------------	0.06	mRNA	IGF2BP1_U	Click
------------------------TACCTCTAAGCACATA----------	0.06	mRNA	IGF2BP1_U	Click
------------TCACCAAACTTGTACCTCTAAGTACATATG--------	0.06	mRNA	IGF2BP1_U	Click
------------TCACCAAATTTGCACCTCTAAG----------------	0.06	mRNA	IGF2BP1_U	Click

Figure 11: "Transcript View" of the location of AGO2/EIF2C CLIP reads along the transcript with Genbank accession NM_182649, which is the human proliferating cell nuclear antigen (PCNA) transcript variant 2. The coding region (CDS), 5′ and 3′ UTR regions are represented as turquoise-colored boxes. The density of reads from the selected samples (AGO2/EIF2C CLIP performed in miR-7-transfected HEK293 cells and IGF2BP1 CLIP in HEK293 cells) along the entire transcript is shown in blue. The user can select transcript regions and visualize the detailed alignment of reads to these regions.



CLIPZ

Mr. Search: CDKN1B [redacted]

Search results for "CDKN1B"

Showing at most 100 results for each category

Transcripts

Number	Transcript Name	Show Alignments	Organism

Genes

Number	Symbol	Description	Organism	Show corresponding Accessions
1	CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)	Homo sapiens	Accessions
2	Cdkn1b	cyclin-dependent kinase inhibitor 1B	Mus musculus	Accessions

Transcripts

Number	Transcript Name	Description	Show Alignments	Organism
1	AB451423	AB451423 Homo sapiens CDKN1B mRNA for cyclin-dependent kinase inhibitor 1B, partial cds, clone: FLJ08132AAAF.	Click	Homo sapiens
2	AF247551	AF247551 Homo sapiens cyclin-dependent kinase inhibitor p27kip1 mRNA, complete cds.	Click	Homo sapiens
3	AJ616234	AJ616234 Homo sapiens partial mRNA for cyclin-dependent kinase inhibitor 1B (CDKN1B gene).	Click	Homo sapiens
4	AK298335	AK298335 Homo sapiens cDNA FLJ51923 complete cds, highly similar to Cyclin-dependent kinase inhibitor 1B.	Click	Homo sapiens
5	AK312461	AK312461 Homo sapiens cDNA, FLJ92816, Homo sapiens cyclin-dependent kinase inhibitor 1B (p27, Kip1)(CDKN1B), mRNA.	Click	Homo sapiens
6	AY004255	AY004255 Homo sapiens cdk inhibitor p27KIP1 mRNA, complete cds.	Click	Homo sapiens
7	BC001971	BC001971 Homo sapiens cyclin-dependent kinase inhibitor 1B (p27, Kip1), mRNA (cDNA clone MGC:5304 IMAGE:3458141), complete cds.	Click	Homo sapiens
8	BT019554	BT019554 Homo sapiens cyclin-dependent kinase inhibitor 1B (p27, Kip1) mRNA, complete cds.	Click	Homo sapiens
9	CR457399	CR457399 Homo sapiens full open reading frame cDNA clone RZPDo834F0810D for gene CDKN1B, cyclin-dependent kinase inhibitor 1B (p27, Kip1); complete cds, incl. stopcodon.	Click	Homo sapiens
10	CR592928	CR592928 full-length cDNA clone CS0DI068YG08 of Placenta Cot 25-normalized of Homo sapiens (human).	Click	Homo sapiens
11	NM_004064	NM_004064 Homo sapiens cyclin-dependent kinase inhibitor 1B (p27, Kip1) (CDKN1B), mRNA.	Click	Homo sapiens

Figure 12: The Search tool can be used to retrieve transcripts by Genbank accession number, gene name or symbol. Binding sites in the transcript of interest are then shown through the "Transcript view".

3.3.2.2 *Search tool*

Another common question is whether any binding sites are known for specific transcripts or genes that a user may be studying. To be able to answer this question we implemented a search tool that allows the user to retrieve from the database a *gene name* or *symbol*, select an accession number associated with it and access the binding site information associated with the transcript in our database (see Figure 12).

3.3.2.3 *miRNA-specific tools*

Because the Argonaute/EIF2C proteins that are part of the RNA-induced silencing complex have been a major focus of the CLIP studies performed so far, we integrated in our server a set of tools that enable the user to explore the identity, abundance and predicted targets of the miRNAs present that were isolated in the CLIP samples. These tools have been described extensively in [71].

3.3.2.4 *Motif enrichment tool*

Finally, one of the main reasons for performing CLIP studies is to determine the sequence specificity of a protein of interest. How a multi-domain RBP contacts RNAs is a challenging question that most likely requires complex computational as well as experimental analyses. However, to provide some preliminary insights we implemented a tool that identifies sequence motifs (defined as n-mers) that are over-represented in an input file (which could contain for instance the most abundant 1000 clusters obtained in an experiment) compared to randomized sequences with the same mono/di-nucleotide composition. We have previously used this tool to show that the motifs that are most over-represented in the clusters from Argonaute/EIF2C PAR-CLIP experiments correspond to the reverse complements of the 5′ end ("seed" region) of the most abundant miRNAs in the cell [67].

3.4 DISCUSSION

Deciphering the post-transcriptional regulatory code that is implemented by regulatory RNAs and RBPs is a problem of great interest [176, 87, 48, 25, 198, 67, 96]. The bottleneck in characterizing RBP binding sites is no longer the availability of an experimental approach, but rather the efficient analysis of the large volumes of data that result from such experiments. Here we present a software system that we developed to analyze data resulting from CLIP experiments. With this system we constructed a database of RBP binding sites that were determined through CLIP and deep sequencing. Our system provides several views of the data, from the level of sequence reads to that of a whole genome browser. Transcript regions with the highest abundance in the CLIP data or that exhibit the highest number of crosslinking events can be easily extracted for further analyses. Both the database and the analysis environment can be easily extended. Registered users can expand the database by submitting their own sequence data sets, the repertoire of organisms can be expanded to include additional species for which a genome assembly is available, and the genome assemblies and transcript databases that are used in the analysis pipeline can be updated as necessary. In the future, we will continue to develop the platform in order to accommodate developments in the sequencing technologies. We expect for instance that the increase in sample size and sequence read length will require the use of heuristic algorithms for mapping short reads to the genome. Such algorithms are in fact already available [117, 118, 107, 192] and will only require one to write adapter programs to interface these programs with the database that stores the alignments. Thus, CLIPZ can eliminate many bottlenecks in the computational analysis of CLIP data and can form the basis for a repository of binding site data for RNA-binding proteins.

FUNDING:

This was supported by the ProDoc program of the Swiss National Science Foundation [Grants PDAMP3_127218 and PDFMP3_123123] and by the Swiss Institute of Bioinformatics.

3.5 ACKNOWLEGEMENTS

We are grateful to Lukasz Jaskiewicz, Shivendra Kishore and the other members of the Zavolan group, to Markus Hafner, Manuel Ascano, and Markus Landthaler of the Tuschl group for providing input and feedback on individual tools, and to Lukas Burger and Jean Hausser for critical comments on the manuscript.

3.5.0.5 *Conflict of interest statement.*

None declared.

4

A QUANTITATIVE ANALYSIS OF CLIP METHODS FOR IDENTIFYING BINDING SITES OF RNA-BINDING PROTEINS

The work presented in this chapter was originally published in Nature Methods [95]

ABSTRACT

Crosslinking and immunoprecipitation (CLIP) is increasingly used to map transcriptome-wide binding sites of RNA-binding proteins (RBPs). To accurately infer RBP binding sites, we developed a method for CLIP data analysis and applied it to compare 254 nm CLIP with PAR-CLIP, which involves crosslinking of photoreactive nucleotides with 365 nm UV light. We found only small differences in the accuracy of these methods in identifying binding sites for HuR, a protein with a low-complexity (U-rich) binding motif and for Argonaute 2, which has a complex binding specificity. We further show that crosslink-induced mutations lead to single-nucleotide resolution not only for PAR-CLIP but also for CLIP. Our results confirm the expectation of the original CLIP publications that RNA-binding proteins do not protect sufficiently their sites under the denaturing conditions of the lysis buffer used during the CLIP procedure, and we show that extensive digestion with sequence-specific ribonucleases strongly biases the set of recovered binding sites. We finally show that this bias can be substantially reduced by milder nuclease digestion conditions.

4.1 INTRODUCTION

RNA-binding proteins (RBPs) are involved in a wide range of processes, from developmental transitions to stress response. Transcriptome-wide identification of RBP binding sites with high-throughput techniques have shown that one regulator typically has hundreds/thousands of targets [19, 37, 106, 137] which tend to be functionally related [85]. Because RBP binding site occupancy depends on the availability of and the crosstalks between regulators [16, 82, 92, 83], substantial efforts are required to characterize their context-dependent biological function.

CLIP, the method of choice for RBP binding site identification, involves crosslinking the protein of interest to target RNAs, immunoprecipitation of the RNA-bound protein and sequencing of the RBP-bound RNA fragments [176]. A few variants, known as HITS-CLIP [120], PAR-CLIP [67] and iCLIP [96], have been proposed. To determine whether differences between protocols are reflected in the set of identified target sites, we started from PAR-CLIP and modified individually those steps that are most likely to bias identification of binding sites: crosslinking and ribonuclease digestion. We focused on two proteins whose binding specificity is well understood and whose binding motifs strongly differ in complexity. The first is HuR, a member of the embryonic lethal abnormal vision (ELAV) family of proteins [6] which binds low-complexity U-rich elements [147]. The second is Argonaute 2 (Ago2 or EIF2C2), a member of the Argonaute family which is guided by miRNAs to induce target gene silencing (reviewed by Ender and Meister [36]). Ago2 has a very complex binding specificity, defined by the entire set of expressed miRNAs.

We performed duplicate experiments for each variant of the CLIP protocol and each protein. We assessed the accuracy of a method both by the reproducibility of results of replicate experiments and through an independent measure. For HuR, we used as independent measures the affinity of the isolated sites for HuR, estimated based on the RNAcompete data [147], and the change in transcript expression after HuR knockdown. For Ago2, the independent measure was the proportion of sites that are complementary to abundantly expressed miRNAs [106, 67]. In our presentation of the results we will refer to the method that employs 4-thiouridine and crosslinking with 365 nm UV light as PAR-CLIP [67], and to the method that does not use photoreactive nucleotides and crosslinks with 254 nm UV as CLIP.

4.2 RESULTS

4.2.1 *Both CLIP and PAR-CLIP reproducibly identify high affinity binding sites for HuR*

We achieved similar crosslinking efficiencies with CLIP and PAR-CLIP by varying the energy up to 1.3 fold (Suppl. Fig. 27). We then identified HuR binding sites in mature mRNAs and computed the enrichment of reads within each site relative to the mRNA expression level as described in Methods. Intersection of the 1000 most enriched sites identified in two replicate experiments yielded 915 sites that were common between CLIP and 862 between PAR-CLIP replicates. Among these, the enrichment in reads is highly reproducible (Suppl. Fig. 28), showing a good correlation (Fig. 13a) with the estimated affinity of the site (computed by averaging the affinity of all 7-mers in the site, as described in the Methods). The correlation changed by less than 0.06 (16%) when we counted only *distinct* reads as opposed to all reads associated with a site, indicating that PCR amplification artifacts during the CLIP procedure were small (Fig. 13a). Contrary to what one may have expected based on PAR-CLIP crosslinking U nucleotides, the U-rich HuR binding sites were not more strongly enriched with PAR-CLIP than with CLIP (Fig. 13).

4.2.2 *Nuclease signature in the CLIP reads*

A second important difference between the published CLIP protocols is that the protein-bound RNAs are fragmented through partial digestion with ribonuclease (RNase) T1/A mix [177, 23], RNase I [188], or micrococcal nuclease (MNase) [201] in HITS-CLIP and through extensive digestion with RNase T1 in PAR-CLIP [67]. We wondered whether at nuclease concentrations that are employed in CLIP experiments, the known sequence specificities of these nucleases are reflected in the nucleotide composition of identified binding sites. We found that with extensive T1 digestion, the correlation between the affinity of individual 7-mers and their enrichment in CLIP binding sites relative to 3′ untranslated regions in which the sites largely reside (Suppl. Fig. 31), is very high for 7-mers devoid of G nucleotides but not for G-containing 7-mers (Fig. 13b-e and Suppl. Fig. 29). The strong preference of RNase T1 to cleave after G nucleotides [169], apparent in the very strong G depletion inside the sequenced reads and the G at the RNA cleavage site (Suppl. Fig. 32a-d), is likely responsible for this effect.

We then substituted the RNase T1 by MNase, which has a 30-fold preference for cleaving 5′ of A or T relative to G or C nucleotides [31] (Suppl. Fig. 32e,f). Although we chose an MNase concentration that yielded 20-50 nucleotide-long RNA fragments, a size range comparable with that obtained with extensive T1 digestion (not shown), the correlation between 7-mer enrichment and affinity (Fig. 13d) was significantly higher in the MNase- compared to the T1-treated samples. The nucleotide composition of the fragments obtained with MNase was also very different (Suppl. Fig. 32e,f), and 7-mers with no As were more strongly enriched relative to 7-mers with comparable affinity but containing at least one A. Replicate MNase-treated samples showed high reproducibility, suggesting that with two different ribonucleases we can obtain *bona fide*, yet quite distinct binding sites (Suppl. Fig. 28).

Finally, we titrated down the concentration of the T1 nuclease from the value specified in the PAR-CLIP protocol to the point where the length of the RNA fragments started to increase (from 30-50 to 50-70 nucleotides, data not shown). The sites recovered with PAR-CLIP with this "mild T1" digestion were only slightly depleted in Gs (Fig. 13e and Suppl. Fig. 29) and the correlation between 7-mer enrichment and affinity (Fig. 13) was higher relative to samples prepared with extensive T1 digestion.

4.2.3 *Crosslink-diagnostic mutations enable high resolution identification of RBP binding sites from both CLIP and PAR-CLIP experiments*

Consistent with the report of Hafner et al. [67] that the position of the crosslink is revealed with PAR-CLIP by a T-to-C mutation presumably introduced as cDNA synthesis progresses over a crosslinked 4-thiouridine, T-to-C substitutions were 8-10 times more frequent than any other mutation in our PAR-CLIP data (Suppl. Fig. 30). This is in stark contrast with the mutational pattern of mRNA-seq reads that we obtained from cells that were either treated with 4-thiouridine or not (Suppl. Fig. 30). In agreement with reports that sequencing through 254 nm UV-crosslinked sites of RBP-RNA interaction induces mutations at the crosslinked position [55], our CLIP data sets also exhibited a mutational bias, distinct from that of PAR-CLIP; T substitutions (to any of A, C, G nucleotides) and deletions were the most frequent mutations in HuR CLIP reads, followed by insertions to either side of T nucleotides (Suppl. Fig. 30). To determine whether these mutations enable high resolution identification of binding sites similar to T-to-C mutations in PAR-CLIP, we extracted 41-nt-long regions centered on the position with the most abundant crosslink-diagnostic mutation (T-to-C in PAR-CLIP, T mutation or deletion in CLIP) in each site and determined the relative location of the ten 7-mers with the highest affinity for HuR [147] in these regions (Fig. 15, Suppl. Fig. 33). We found that the centers of the high affinity 7-mers, probably positioned between the RNA recognition motifs 1 and 2 of HuR [184], are located at very specific distances relative to the most frequently mutated positions. For example in CLIP, the crosslink occurs most frequently at the first T after the non-T base of the TTTATTT, TTTCTTT and TTATTTT 7-mers. The location of the crosslink relative to the TTTTTTT 7-mer is less precise, most likely due to the fact that the protein can be captured at different positions on homogeneous T stretches. Interestingly, the TTTTTTT 7-mer appears more frequently downstream of the position with the most abundant mutations in CLIP compared to PAR-CLIP (52.5% vs. 41.8%, averages between replicate experiments).

The enrichment of reads in a binding site relative to the expression level of the mRNA in which the site resides is a natural measure of the "quality" of the site (its affinity for the RBP). It has been argued however [67], that the number of crosslink-diagnostic mutations in a putative site is also indicative of the site’s "quality", bypassing the need to estimate transcript expression levels to compute enrichment. Consistently, we found that for both CLIP and PAR-CLIP the predicted affinities were comparable between sites extracted based on their enrichment in reads or based on the density of crosslink-diagnostic mutations (defined above) (Fig. 14).

A summary of the HuR binding sites that are obtained with different variants of the CLIP protocol is shown in Supplementary Table 1 and the datasets can be further explored through the CLIPZ server that is developed by Khorshid et al. [90].

4.2.4 *Functional validation of CLIPed targets of HuR*

As a final test of accuracy of various CLIP methods in identifying functional HuR binding sites, we monitored the change in expression of CLIPed targets upon siRNA knockdown of HuR. Because ELAV family members predominantly stabilize and promote translation of target mRNAs [114, 38], we expected that HuR knockdown results in reduced expression of HuR targets. We estimated transcript expression levels with mRNA-seq and used as reference data from cells treated with a mock siRNA directed against the green fluorescent protein (GFP), and data from untreated cells. We found that HuR targets are enriched 1.5-2 fold (Suppl. Fig. 34a) relative to all transcripts within a given starting expression range that are destabilized by HuR knockdown (see also Methods). Interestingly, only for sites whose enrichment was inferred from samples treated with MNase or mild T1 does the strength of destabilization increase with increasing site enrichment, further suggesting that the set of binding sites recovered under these conditions is less biased (Suppl. Fig. 34b).

4.2.5 *High-resolution identification of miRNA target sites with CLIP and PAR-CLIP*

We further investigated whether the various CLIP methods are equally efficient in identifying more complex RBP binding sites as those of Argonaute proteins, which are guided to their targets by small RNAs. Like HuR, Ago2 can be efficiently crosslinked with either 254 nm UV, or with 365 nm UV after 4-thiouridine treatment (Suppl. Fig. 27). Various CLIP variants also yield reproducible sets of Ago2 targets (Suppl. Fig. 36a-c), reproducibility being highest among replicates that were obtained using the same method and then among samples that were prepared with the same nuclease treatment (proportion of sites in common between two samples in the range of 36-65% and 28-54%, respectively among the top 1000 sites, see Suppl. Fig. 36d). Because many studies showed that complementarity to the "seed" region (positions 2-8 from the 5′ end) of miRNAs is most predictive for changes in mRNA levels in response to changes in miRNA concentration (see e.g. Lewis et al. [116]), we used the proportion of binding sites that are complementary to the seed of the most abundantly expressed miRNAs (for details see Methods) as an independent measure of the quality of the target set. We found that this proportion was highest for the sites with the strongest enrichment in reads (Fig. 16), and that ranking sites based on the density of crosslink-diagnostic mutations yielded comparable results (Fig. 16). PAR-CLIP, especially in conjunction with MNase treatment, yielded a higher proportion of miRNA seed-complementary sites than CLIP.

Following our results with HuR and previous work [162, 75] we used T substitutions and deletions as crosslink-diagnostic mutations in Ago2-CLIP. As shown in Fig. 17 (as well as Suppl. Figs. 38 and 39), we found that 254 nm UV-induced mutations occur immediately upstream of the miRNA seed-complementary region, similar to what has been observed with PAR-CLIP [67]. This is not due to a specific nucleotide composition bias in the immediate vicinity of the seed matches (Suppl. Fig. 38h). As T mutations are less abundant in Ago2-CLIP compared to HuR-CLIP (Suppl. Fig. 30) we investigated whether other nucleotides are also targeted with CLIP, indicating the position of the crosslink. We extracted the top 1000 most enriched Ago2 sites, located the position within each site where most mutations of a particular type occurred, and determined the locations of miRNA seed matches with respect to this position. We found that deletions (of any of the 4 nucleotides) and to a lesser extent mutations of nucleotides other than T also occur predominantly immediately upstream of miRNA seed matches (Suppl. Fig. 39). In contrast, insertions were located 7-10 nucleotides upstream of the miRNA seed match (Suppl. Fig. 38g).

4.2.6 *Nuclease signature in the Ago2 CLIP samples*

Because measurements of affinities of miRNA-containing Ago2 for binding sites are not available, we analyzed instead the nucleotide composition of the top 1000 Ago2 target sites from different samples. We found that the frequency of guanosines is lower in CLIP and PAR-CLIP samples prepared with RNase T1, while MNase-treated samples show a clear, though less marked depletion in A's and T's. These results are consistent with the MNase preference for cleavage 5′ of these nucleotides and the milder digestion conditions (Suppl. Fig. 40).

4.3 DISCUSSION

CLIP approaches combined with next generation sequencing have been successfully employed to identify targets of RBPs. However, because no study investigated the relationship between the affinity of the RBP for individual binding sites and the number or enrichment of reads obtained from these sites in CLIP experiments, little is known about how different CLIP protocols compare in the identification of RBP binding sites.

We found that for the two proteins that we studied, HuR and Ago2, comparable RNA yields can be obtained with CLIP and PAR-CLIP by varying the amount of energy applied for crosslinking by a factor of 1.3. The relative RNA yields were however, protein-dependent: for HuR we obtained a ~5-fold higher amount of RNA when we crosslinked with 254 nm UV light, while for Ago2 4-thiouridine-mediated crosslinking at 365 nm yielded ~10-fold more RNA. Hafner et al. [67] made a different comparison, showing that *at the same energy* employed for crosslinking, RNA recovery after crosslinking of IGF2BP1 is 100-1000 fold higher when 4-thiouridine is used. Employing the RNAcompete dataset of estimated affinities of HuR for all possible 7-mers [147], we found that for this U-rich element binding protein, 254 nm UV crosslinking yields a higher correlation between the affinity of sites and their enrichment in CLIP. On the other hand, for Ago2, a protein whose targeting specificity is given by guiding miRNAs, PAR-CLIP yields miRNA profiles that resemble closer the levels of mature miRNAs in total RNA and mRNA binding sites with higher enrichment in miRNA seed-complementary motifs compared to CLIP. Contrary to the expectation that sites that are isolated with PAR-CLIP are enriched in U nucleotides compared to sites that are isolated with CLIP, we found that the U-rich binding sites of HuR are most efficiently isolated with CLIP. We believe that this is due to multiple U's in a binding site being available for crosslinking at 254 nm, whereas PAR-CLIP restricts the possibility of crosslinking to the photoreactive nucleotide analog. At a photoactivatable nucleoside incorporation rate of 1 in 40 (estimated in Hafner et al. [67]) it is unlikely that a single HuR binding site contains more than 1 nucleotide that can be crosslinked with PAR-CLIP.

4.3.1 *Identification of binding sites, read enrichment and diagnostic mutations*

One of the main motivations behind CLIP approaches to RBP binding site identification is that they hold the promise of very high (site/nucleotide) resolution. This is important for inferring the sequence-specificity of a protein, or the determinants of specificity beyond the sequence of the binding site. If the RBP protected its binding sites from nuclease cleavage under the conditions of the experiment, complete digestion of all accessible RNA regions followed by immunoprecipitation would indeed lead to unbiased isolation of binding sites. Practically, the protection conferred by RBPs is likely insufficient (or even absent under strong denaturing conditions), leading to isolation of RNA fragments that are too short to be assigned with any degree of confidence to particular mRNAs when the nuclease digestion is extensive.

Two approaches were previously taken to circumvent this problem. In Hafner et al. [67] RNA digestion, though extensive, was performed with the T1 ribonuclease which cleaves very specifically 3′ of G nucleotides, still allowing isolation of large numbers of functional binding sites for the proteins covered in that study. The drawback of this approach is that the very high nuclease specificity combined with the incomplete protection that the RBP provides to its sites leads to depletion of binding sites that contain G nucleotides within or in their immediate vicinity. Therefore, obtaining accurate CLIP data for proteins with G-rich binding sites or that are composed of multiple domains with very different sequence specificity will be difficult. The second approach, typically taken by HITS-CLIP [120] is to use more controlled nuclease activity. The data from samples with mild MNase and RNase T1 treatment that we presented here indicate that combined with appropriate computational analysis, this approach enables identification of high-affinity binding sites with high resolution and minimal nuclease-specific bias. Careful titration of the nuclease amount is in this case recommended because the interplay between nuclease specificity and concentration, and the nucleotide composition of the regions in which the binding sites for the protein of interest reside will determine what binding sites can be isolated. One of the issues to be dealt with in this approach is separation of binding site-derived reads from non-specifically isolated fragments of abundant RNAs (background). Chi et al. [23] estimated the background based on microarray measurements of mRNA expression and simulation. Here we use instead mRNA-seq data. Correcting for the abundance of individual mRNA species in total RNA had very little effect on the average affinity of the top-ranking sites from extensively digested samples, but led to increased average affinity of binding sites from samples prepared with milder digestion conditions. The improvement was small in the case of HuR sites (Fig. 14), while for Ago2 there was a 10% increase in the fraction of sites (among the top 1000) with a seed match (Fig. 16) relative to the ranking based on coverage by CLIP reads alone. Thus, although simple ranking by the coverage by reads or density of crosslink-diagnostic mutations enables identification of binding sites, we do recommend estimating the mRNA expression level in the cell type of interest with a method such as mRNA-seq and ranking the sites by enrichment in CLIP relative to mRNA-seq. This correction should be especially important when the antibody specificity is not very high.

Crosslink-diagnostic mutations are apparent with both CLIP and PAR-CLIP. In 4-thiouridine-treated samples that were crosslinked with 365 nm UV these mutations were predominantly T-to-C substitutions in the cDNAs, followed by T deletions and insertions to either side of T nucleotides. 254 nm UV crosslinking induces mutations that are 2-4 fold less abundant and are more complex. The preferred location of mutations was immediately upstream of the miRNA seed-binding region for Ago2 and at the center of high affinity motifs, possibly positioned between the first and second RNA-recognition domains of the protein [184], for HuR. The specific location of mutations within binding sites should be particularly useful for the inference of binding sites of RBPs with a complex binding specificity that cannot accurately be described by a weight matrix.

Based on the studies available to date, two approaches thus appear most promising for accurate, high resolution identification of binding sites. One involves partial digestion with a relatively unspecific nuclease, peak finding to locate the binding sites, ideally taking into account relative abundance of mRNAs, and analysis of mutations to locate the crosslinked residues. One drawback is that multiple proteins may be crosslinked to long RNA fragments, which consequently will not migrate at the expected size in the protein gel, and will likely be selected against in the CLIP procedure. The second approach is to identify the position of the RBP-RNA crosslink already during sample preparation, as done in iCLIP [96], a modification of CLIP that attempts to take advantage of the tendency of reverse transcriptase to prematurely terminate at crosslinked nucleotides. Because this protocol cannot be implemented by simply varying a step in PAR-CLIP, we did not include iCLIP into our study.

Studies of various RBPs with CLIP have already revealed extensive and complex networks that regulate mRNA processing, traffic, stability and translation. Because CLIP approaches will likely become common in the future, we used two proteins already shown to be involved in crosstalks in post-transcriptional gene regulation to study the effect of various choices that are made in CLIP protocols on the accuracy of the data. Our findings will enable the design of improved protocols to further uncover post-transcriptional regulatory networks.

4.4 ACKNOWLEDGMENTS

We thank G. Meister for Ago2 antibody, C. Rammelt for initial help with PAR-CLIP, I. Nissen and C. Beisel for support with deep sequencing, J. Krol for help with miRNA quantification, B. Dimitriades for help with cell culture, W. Keller and W. Filipowicz for critical comments and suggestions. We also thank the members of the Zavolan lab for valuable discussions. This work was supported by grants from the Swiss National Science Foundation (#31003A_127307, PDAM_3_127218 and PDFMP3_123123), the Swiss initiative in systems biology (CellPlasticity) and the Swiss Cancer League (#KFS 02477-08-2009) and by EMBO fellowship ALTF 184-2009 to S.K.

4.5 AUTHOR CONTRIBUTIONS

S.K. designed and performed the experiments, L.J. designed and performed the experiments and wrote the paper, L.B. analyzed the data, J.H. analyzed the data, M.K. developed the annotation tools and analyzed the data, M.Z. designed and supervised the experiments, analyzed the data and wrote the paper.

4.6 METHODS

4.6.1 *CLIP and PAR-CLIP*

HuR was immunoprecipitated with anti-HuR (N-16) antibody from Santa Cruz Biotechnology; anti-Ago2 antibody (11A9) was a gift from G. Meister [151]. Crosslinking, immunoprecipitation and library preparation were carried out according to the protocol established by Hafner et al. [67] with the following modifications. For crosslinking at 254 nm, cells were irradiated on ice using Stratalinker 2400 (Stratagene) twice at 0.1 J/cm^2 with 1 min break. In the MNase-treated samples for mild initial cleavage step to break down large protein complexes we used 50 U/ml (instead of 1000 U/ml, as in the original protocol) of RNase T1 (Fermentas) and the second step of nuclease digestion was carried out on the beads exactly as in PAR-CLIP protocol, but with MNase (New England Biolabs, 0.2 gel U/µl final concentration) instead of RNase T1. Since MNase activity is very sensitive to the buffer conditions, MNase digestion was carried out in the buffer supplied with the enzyme and incubated at 37 °C for 5 min. For mild RNase T1 sample we used 5 U/ml for the initial cleavage in the lysate and 20 U/µl for the second step on the beads. Library amplification step was performed with minimal number of PCR cycles that still enabled us to see the DNA product on the agarose gel in the pilot PCR. With the exception of Ago2-CLIP B, for which we used 20 cycles, all samples were amplified for 16-18 PCR cycles.

4.6.2 *mRNA-Seq*

Total RNA from HEK293 cells was isolated with TRI Reagent (Sigma). Libraries were prepared using mRNA-Seq Prep Kit (Illumina).

4.6.3 *miRNA profiling by qRT-PCR*

As suggested by Hafner et al. [67], we used the CLIP data to identify the miRNAs that were most abundant in the crosslinked Ago2-containing complexes. This is because UV light crosslinks Ago2 not only to its targets but also to the guiding miRNAs. The relative abundance of Ago2-bound miRNAs correlated strongest between replicate samples, followed by samples that underwent the same nuclease treatment (Suppl. Fig. 37a). To further determine whether the CLIP/PAR-CLIP-based miRNA profiles reflect the miRNA expression levels in total RNA, we selected nine miRNAs that were relatively abundant, but still covered a hundred fold range of expression (Suppl. Fig. 37b) in different samples and we measured their expression in total RNA by quantitative RT-PCR. Total RNA from HEK293 cells was isolated with TRI Reagent (Sigma). Reverse transcription of mature miRNAs and quantification was performed with TaqMan miRNA assays (Applied Biosystems) as described in Krol et al. [97] The following miRNAs were assayed: let-7a, let-7f, miR-16, miR-19b, miR-27b, miR-30c, miR-92a, miR-301 and miR-424. qPCR was done using a Corbett Rotor Gene 3000 system. The threshold cycle (Ct) values were determined using default threshold settings. We found that although the abundance of these miRNAs was tightly correlated between replicate CLIP samples, their expression levels in total RNA correlated best (average correlation over the two replicate experiments of 0.65) with the levels inferred from the PAR-CLIP samples (Suppl. Fig. 37). To determine the proportion of Ago2 sites that were complementary to the most abundantly expressed miRNAs we therefore used the miRNA profiles determined based on PAR-CLIP MNase experiments.

4.6.4 *siRNA transfections*

HuR knockdown analysis was performed with siRNA against HuR (sc-35619, Santa Cruz Biotechnology). In brief, HEK293 cells were grown in 6-well plates and each well was transfected with 150 pmoles of siRNAs with 10 µl of Nanofectin siRNA reagent (PAA Laboratories) as per manufacturer's instructions. The medium was changed after 5 hours of transfection and the cells were harvested for protein and RNA analysis after 96 hours of siRNA treatment. The expression of HuR estimated based on mRNA sequencing, decreased to about 30% upon siRNA knockdown, consistent with the change we observed on the Western blot (Suppl. Fig. 35).

4.6.5 *Estimation of transcript expression*

We used mRNA-annotated reads that mapped uniquely to genic regions of the genome to estimate transcript expression based on mRNA-seq data. For each gene, we selected from Genbank one representative transcript per Entrez gene. This was the longest transcript associated with the gene, preference being given to transcripts that could be mapped to the genome (hg18 assembly version from the University of California, Santa Cruz, [1]), and among these to transcripts that are part of the Refseq database of NCBI [2]. For mapping sequenced reads to transcripts and to the genome assembly, and for functional annotation we used a procedure that we described before [14, 67, 90]. In subsequent analyses we used mRNA-annotated reads that mapped uniquely to the genome. We defined the transcript expression level as the density of reads per nucleotide assuming a standard total number of one million reads in the sample.

[1] http://hgdownload.cse.ucsc.edu/downloads.html

[2] http://www.ncbi.nlm.nih.gov/refseq/

4.6.6 *From reads to binding sites*

Various analysis methods have been previously proposed for CLIP data [23, 67], taking advantage of the specific experimental design and the peculiarities of the data sets. In this study we designed the following method that we applied uniformly to all data sets in order to identify RBP binding sites. We aimed to identify reliable CLIPed sites in transcripts that are expressed at a sufficiently high level for which we can accurately compute the enrichment relative to total mRNA expression. Because the proteins that we studied bind regulatory sites located primarily in 3′ UTRs and because our protocol isolates cytoplasmic RNA, we computationally analyzed sequences of mature mRNAs.

We determined the expression of individual transcripts in the total RNA by mRNA sequencing and we found that the density of sequenced reads within transcripts is bimodal, with a low density peak probably corresponding to transcripts of low abundance and transcripts to which reads were spuriously mapped, and a high density peak of about 10 reads per kb of transcript per million reads in the library (not shown). By fitting a two-component gaussian mixture to the data (using the R package mclust [43]), we isolated the transcripts with reliable expression as given by mRNA sequencing. We found that the fitted distributions were very similar between replicate mRNA sequencing data sets and that the gaussian mixture fitting yielded between 12′517 and 13′493 expressed transcripts per mRNA-seq data set, with 11′564 being common to all four mRNA-seq data sets.

Because the CLIP protocols involve size selection of RNA fragments and we isolated fragments in the range of 20-70 nucleotides from the gel (sequencing up to 36-38 nucleotides), we chose 40 nucleotides as the length of the binding regions (peaks) that we set to uncover from the CLIP data. Although the CLIP tags were of varying lengths, the average coverage (number of reads that overlap a position) per nucleotide in peaks showed a clear quantization pattern, particularly so in the extensively digested T1 samples, with a clear separation between peaks with many reads and peaks with 1-2 or fewer reads. To remove these very low-coverage peaks in an automated way, we again used a gaussian mixture model approach. Since in this case the component with larger mean generally had a much larger variance than the component with the smaller mean, we further removed the small fraction of regions that were classified as belonging to the component with higher mean yet whose coverage was smaller than the mean of the component with lower mean. Finally, we selected binding regions located in transcripts that we considered expressed and we computed the ratio between their coverage and the average coverage in a 40 nucleotide window of the corresponding transcript as given by mRNA-seq.

To evaluate the robustness of our results we performed similar analyses on sites that we extracted based on other criteria such as the enrichment of reads with respect to mRNA expression setting the counts for each distinct tag to one (to mitigate potential amplification biases), the density of crosslinking-diagnostic mutations either in a binding region in a mature mRNA or in the genome, or the density of reads in a binding region defined on the genome. In the latter case we tested both sites extracted based only on uniquely-mapped mRNA-annotated reads as well as sites extracted based on uniquely and multiply mapped reads that were annotated as mRNA or repeat. The results are robust with respect to these different approaches to binding site identification.

The library preparation procedure for most of the currently available next generation sequencing technologies requires that the cDNAs are amplified by PCR. This step has the potential to hinder accurate quantification of read abundance because some sequences are more efficiently amplified during PCR than others. Various approaches have been employed to minimize this bias. On the experimental side, ligation of random barcodes to individual sequences in the initial sequence pool was employed in previous CLIP studies [120, 96]. On the computational side, collapsing identical reads and counting only distinct reads was previously tried in the attempt to minimize amplification bias. Here we did not use random barcoding during sample processing, but we limited the number of PCR amplification cycles to 16-18. The correlation between the predicted affinities of sites and their enrichment, calculated either by counting all reads or only the distinct reads, was similar, suggesting that amplification bias was not substantial in our experiments. Nonetheless, improving the quantification accuracy at the experimental level through methods like random barcoding should further improve the identification of high-affinity binding sites.

4.6.7 *HuR knockdown analysis*

For both the HuR siRNA knockdown and the control (siRNA directed against the green fluorescence protein (GFP)) RNA-seq sample, transcript expression levels were computed as described above. To eliminate potential biases due to differences in the shape of the distributions of expression levels, the two siRNA-treated samples were quantile-normalized with the normalizeQuantile function of the limma R package [163] [3]. We then binned all expressed transcripts into five bins of equal size based on their expression in the GFP siRNA-treated sample and, for each bin separately, we computed the mean and variance of the $\log_2$ fold-change in transcript expression level in the HuR-siRNA experiment and converted $\log_2$ fold-changes into Z-values. We

[3] http://www.R-project.org

identified the transcripts whose expression was down-regulated in the HuR siRNA-treated sample relative to the GFP siRNA-treated sample at varying negative Z-value cutoffs, and determined the enrichment of HuR targets, defined as the transcripts to which at least one of the top 5000 peaks mapped, relative to all transcripts within each expression bin. We used a cut-off Z-value of -1 (resulting to a total of 1946 down-regulated transcripts) in order to have a sufficiently large number of down-regulated transcripts in each expression bin to calculate stable enrichments. We obtained qualitatively similar results with a more conservative cut-off of -2. In order to determine whether the fold-change of transcripts was related to the enrichment of their sites in CLIP reads, we we successively selected, from each CLIP sample, the top 1000, 1001-2000, 2001-3000, ..., 4001-5000 binding sites and determined the distribution of fold-changes of the transcripts from which these subsets of binding sites originated. Additionally, we determined the fold-changes of all expressed transcripts to which none of the top 5000 sites mapped. We further established that choosing as control mRNA-seq data from untreated samples as opposed to GFP siRNA-treated samples led to similar conclusions (not shown).

4.6.8 *Mutation Analysis*

To analyze the mutational signature of the CLIP and mRNA-seq libraries, we considered, for each library, all mRNA-annotated reads that mapped uniquely to the genome. We determined the frequency of each possible substitution (from any nucleotide X to any nucleotide Y), the frequency of deletions in any of the four nucleotides as well as the frequency of insertions. In the case of insertions, we distinguished between the identity of the inserted nucleotide and the identities of the nucleotide to the right and left of the insertion. Mutation frequencies were defined as the number of occurrences of a particular mutation divided by the total number of nucleotides in all mRNA-annotated tags. As the mutation frequency depends on the number of errors that are allowed for the mapping of the reads to the genome, the mutation frequencies that we computed are likely underestimates of the true mutation frequencies and should only be interpreted in relative terms.

4.6.9 *Affinities of CLIPed regions*

As an estimate of the affinity of HuR to each possible heptameric sequence, we used the enrichment scores from previously published RNAcompete experiments [147]. For each heptamer, we averaged the enrichment scores of the two replicate experiments described in the study. To determine the affinity of the 40nt-long binding regions, we averaged the heptamer scores[147] over all subsequences of length 7. The heptamer enrichments in CLIPed regions relative to 3′ UTRs were calculated as the ratio of the heptamer frequencies in the top 1000 binding regions and their frequencies in 3′ UTRs, which were determined by counting all 7-mers in the 3′ UTR of our set of representative transcripts (see above). To avoid spuriously high enrichments due to low counts, we first rescaled the counts for each 7-mer in 3′ UTRs such that the total number of 7-mers in 3′ UTRs equalled the total number of 7-mers in the CLIPed regions. We then added to the counts in 3′ UTRs as well as in CLIPed regions a pseudocount of 10. The enrichment of each 7-mer was then defined as the ratio of these counts in CLIPed regions and in 3′ UTRs.

4.6.10 *Extraction of crosslink-centered regions*

We started from binding sites identified on the basis of their enrichment, as described above, and tabulated the frequency of various types of mutations within these binding sites. We then identified the position of the most abundant mutation of a specific type and extracted a symmetrical region around this position. We called these crosslink-centered regions and used them to investigate whether various types of sequence motifs occur at a specific position with respect to the most abundant mutation in the site.

4.6.11 *Identification of the ten highest expressed miRNA families*

We obtained the number of sequence reads that mapped to each mature miRNA with the annotation pipeline described in Khorshid et al. [90] The counts were then aggregated by miRNA families, which were defined by the subsequence at positions 2-8 of the mature miRNAs. The ten most expressed miRNA families for a given experiment type (Ago2 CLIP, PAR-CLIP and CLIP-MNase) were determined by averaging counts from the two replicates. Unless specified, the ten most expressed miRNA families from the Ago CLIP-MNase samples were used in the analysis of targets.

4.6.12 *Fraction of sites with a match to one of the top most expressed miRNA families*

We ranked sites based on various measures: average coverage of a nucleotide by reads obtained in a CLIP experiment, enrichment in sequence reads relative to the mRNA expression, and density of crosslinking-diagnostic mutations. When then took bins of 1000 sites, that is the sites with ranks 1-1000, 1001-2000, 2001-3000, 3001-4000 and 4001-5000, and we determined the fraction of sites within each bin that matched the seed of at least one member of the ten most expressed miRNA families. A seed match was defined as a 7-mer motif complementary to positions 2-8 of a miRNA. The standard errors were computed from the observed fraction q of sites with a match and the sample size n=1000 using the relation $\sqrt{\frac{q(1-q)}{n}}$.

4.6.13 *Location of HuR and miRNA-complementary motifs with respect to the position of crosslink*

We started with the 1000 sites with highest enrichment in reads and extracted the 41 nucleotide-long sequences centered on the location of the most frequent crosslink-diagnostic mutation. In the case of HuR, we determined, separately for each of the ten 7-mers with highest affinity separately, the frequency of matches relative to the location of the crosslink. A match was anchored at the center of each 7-mer. In the case of Ago2, we searched for occurrences of matches to the seeds of the ten most expressed miRNA families. The intensities in the Ago2 heatmaps correspond to the number of crosslink-centered region that match the positions 2-8 of the miRNA families indicated on the y-axis. In Suppl. Fig. 38, the sequence logo was constructed based on the 10 nucleotides upstream of matches to the seed of one of the ten most expressed miRNA families in the 1000 sites with highest enrichment in reads. In case several seed matches could be found in the same site, the contribution from the regions upstream of the seed matches were divided by the number of found seed matches, so that the contribution of each of the top 1000 sites to the sequence logo was equal.

4.6.14 *Correlation of enrichment in replicate samples*

In each sample, we focused on the 1000 sites that were most enriched in reads relative to the expression of the mRNAs in which they occurred. We determined the fraction of sites that overlapped by at least 1 nucleotide across replicate experiments. For these sites, we plotted the enrichment in replicate A vs the enrichment in replicate B.

4.6.15 *Reproducibility of miRNA expression*

For all 6 Ago CLIP libraries, we combined the reads mapping to mature miRNAs by miRNA families, defined as sets of miRNAs with identical 2-8 positions. Scatters shown in Suppl. Fig. 37 show the reproducibility of the $\log_{10}$ read counts, where each dot corresponds to a miRNA family expressed in both samples, and the red dots represent miRNA families that were measured by qPCR. We also report the correlation coefficient between the $\log_{10}$ read counts across replicates for all miRNA families expressed in both samples.

4.6.16 *Nucleotide composition of Ago2 sites*

For each of the 6 Ago CLIP libraries, we obtained the 1000 sites most enriched in sequence reads and we determined the proportion of A, C, G, and T nucleotides in these sites.

4.6.17 *Sequence composition around the 5′ and the 3′ ends of reads obtained in different samples*

Position-wise nucleotide frequencies along the sequence reads were determined based on all the sequence reads that mapped to the top 1000 sites (according to enrichment with respect to mRNA expression). Sequence logos were drawn using the R package seqLogo [12] [4].

4.6.18 *Observed and expected distribution of reads among 5′ UTR, CDS and 3′ UTR regions of transcripts*

Reads were annotated as described in Khorshid et al. [90] Based on the mappings of reads to transcripts with annotated coding region we assigned reads to 5′ UTR, CDS and 3′ UTR regions. When a read had multiple mappings, the count of the read was distributed equally between the alternative loci. The expected distribution was computed assuming that the read could have come from any of the 5′ UTR, CDS or 3′ UTR regions of the transcript(s) to which it mapped, with relative probabilities given by the relative length of those transcript regions.

[4] http://www.bioconductor.org/packages/release/bioc/html/seqLogo.html

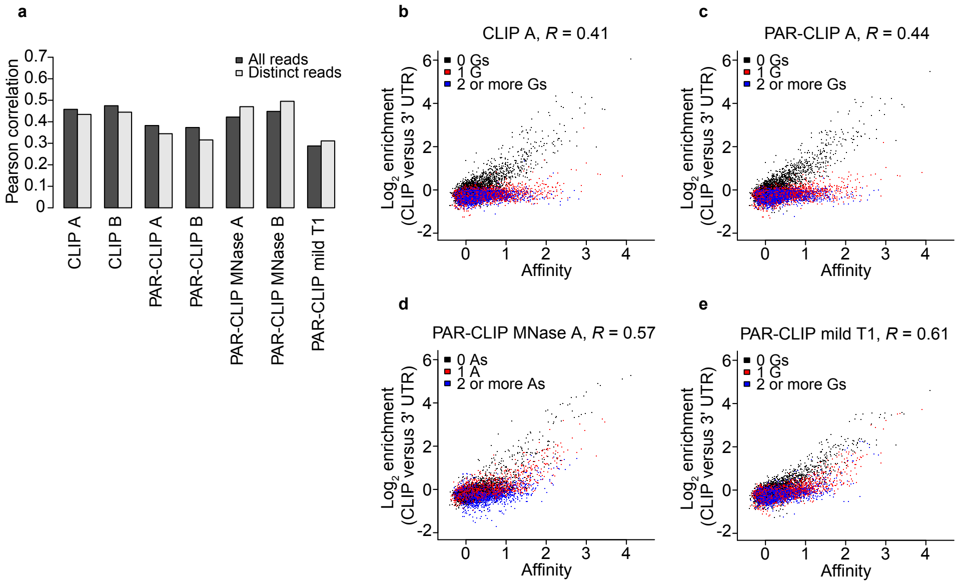


Figure 13: Pearson correlation coefficients between the enrichment in reads (relative to mRNA abundance) of HuR binding sites identified by various CLIP and PAR-CLIP variants and their predicted affinity. To investigate a potential amplification bias, we show the same correlations for the top binding sites when only distinct reads are counted (light grey bars). Correlation between the estimated affinity of a 7-mer motif and its enrichment in CLIP (b), PAR-CLIP (c), PAR-CLIP MNase (d) and PAR-CLIP mild T1 (e) binding sites relative to 3′ UTRs.

FIGURES

In this section, figures of this study are illustrated

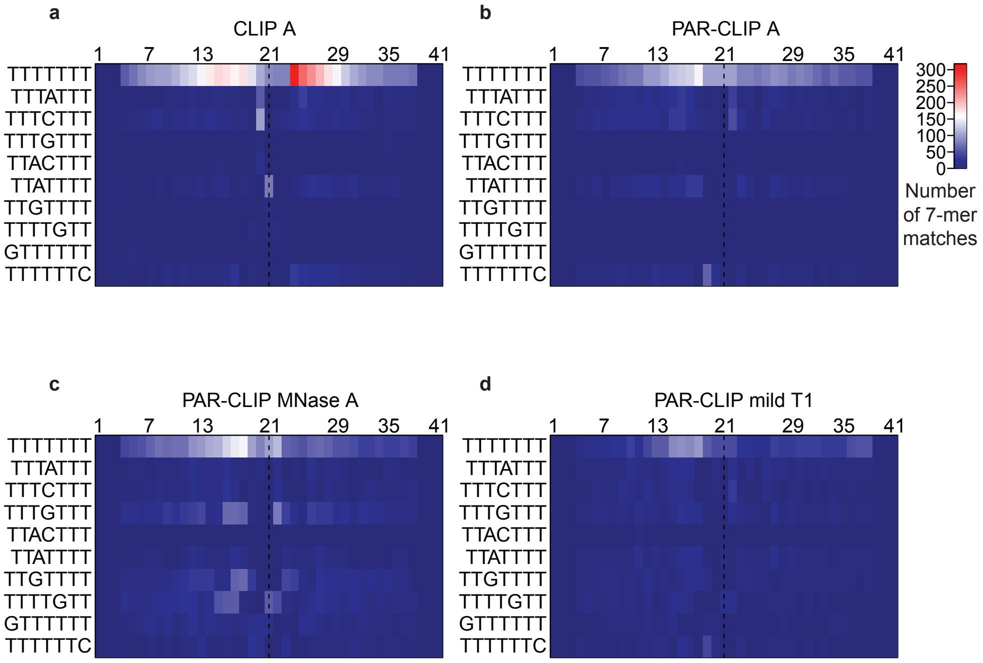


Figure 14: Distribution of predicted affinities of HuR binding sites that are isolated based on different measures. These measures were (a) enrichment relative to the abundance of the mRNA in the total RNA, (b) coverage of the binding site by reads, and (c) density of crosslink-diagnostic mutations in a given site. For each measure and each sample, the sites were sorted and then divided into non-overlapping bins of 1000 sites, which are shown from the left-to-right for each individual experiment. Error bars indicate standard error of the mean

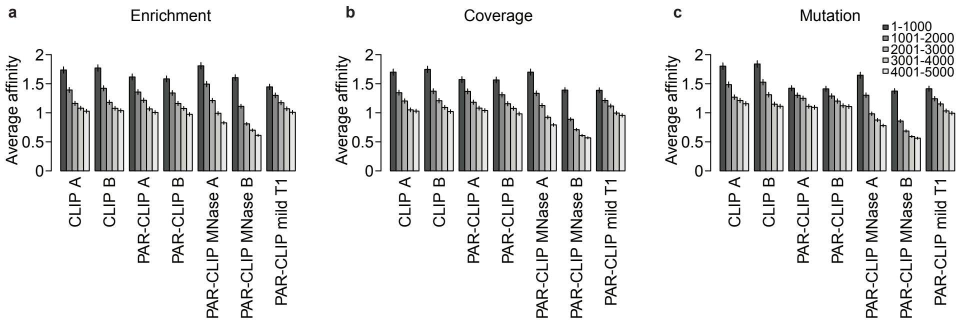


Figure 15: Location of the ten 7-mers with highest affinity for HuR relative to the crosslink site. Anchoring each 7mer at its central position, the frequency of 7-mer matches as a function of the distance to the crosslink site was determined for CLIP (a), PAR-CLIP (b), PAR-CLIP MNase (c) and PAR-CLIP mild T1 (d). The position of the predominant mutation (T deletion or mutation to G/A/C in CLIP and T-to-C in PAR-CLIP) is indicated by a dashed line.

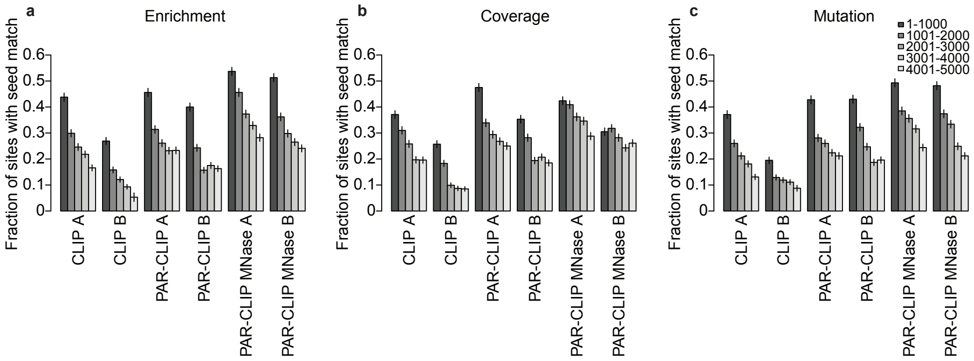


Figure 16: Proportion of Ago2 binding sites matching the seed regions of abundantly expressed miRNAs. For the top 1000, 1001-2000, etc. Ago2 binding sites identified based on enrichment (a), coverage by reads (b) or density of crosslink-diagnostic mutations (c), we determined the fraction of sites that are complementary to one of the ten most abundant miRNA seed families according to the miRNA profile of the MNase-treated PAR-CLIP samples. Error bars represent standard errors on the fraction of binding sites with seed match.

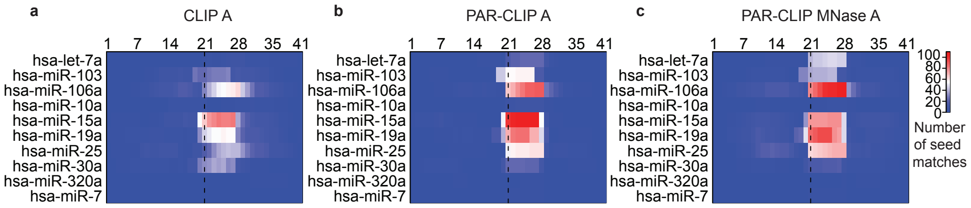


Figure 17: Location of miRNA seed-complementary regions relative to the crosslink-diagnostic mutation. Considering the ten most abundant miRNA families, we determined the location of the miRNA seed matches relative to the position of the crosslink-diagnostic mutation (which is in the center of the 41-nucleotide-long region) in the 1000 most enriched Ago2 sites in CLIP (a), PAR-CLIP (b) and PAR-CLIP MNase (c).

Part III

MATHEMATICAL MODELING

5 MIRZA: A BIOPHYSICAL MODEL FOR INFERRING MICRORNA-TARGET SITE INTERACTIONS FROM ARGONAUTE CROSSLINKING AND IMMUNOPRECIPITATION DATA

The work presented in this chapter was originally published in Nature Methods [91]

ABSTRACT

We introduce a biophysical model of miRNA-target interaction and infer its parameters from Argonaute 2 crosslinking and immunoprecipitation data. Combining this model with miRNA transfection data, we show that a substantial fraction of miRNA target sites are non-canonical, and that predicted target site affinity correlates well with the extent of target destabilization. Our model provides a rigorous biophysical approach to miRNA target identification beyond ad hoc miRNA-seed based methods.

5.1 MAIN

miRNAs are a large class of regulators of gene expression, post-transcriptionally modulating the stability of mRNA targets and their rate of translation into proteins. Although in mammals, 7-8 nucleotides of perfect complementarity between the miRNA 5′ end and the target mRNA is frequently sufficient to elicit a response (typically measured in terms of mRNA degradation [10]), many such ′miRNA seed′-matching sites have no apparent effect. Thus, current target prediction methods additionally make use of conservation and sequence context information to reduce false positive predictions [44, 46]. ′Non-canonical′ sites, that are not perfectly complementary to the miRNA seed region yet are effective in down-regulating gene expression, have also been described [109, 191, 182, 102]. However, they are considered rare, and the currently most-accurate prediction methods [10, 2] do not attempt to identify them.

Recently developed methods for Argonaute protein crosslinking and immunoprecipitation (Ago-CLIP) [23, 201, 67, 95] enable experimental identification of miRNA binding sites transcriptome-wide. While this provides the opportunity to investigate in detail the principles and consequences of miRNA-mRNA target interaction, Ago-CLIP on its own does not identify which miRNA guided Ago to each binding site, or the structure of the miRNA-target site hybrid. Here we introduce a rigorous biophysical model of miRNA-target interaction and infer its energy parameters from Ago-CLIP data. The model (which we called MIRZA and is described in detail in the Methods) includes, besides parameters associated with base-pairs and loops, specific miRNA position-dependent energy parameters that reflect the constraints imposed by the Argonaute protein on miRNA-mRNA interaction. Figure 18A illustrates how MIRZA calculates the energy of a possible miRNA-mRNA target hybrid in terms of its 27 energy parameters.

We infer MIRZA's parameters by maximizing the ratio $R(D)$ of the probability to obtain the data set $D = (m_1, m_2, \ldots, m_n)$ of CLIPed mRNA fragments by immunoprecipitation with Ago as opposed to randomly sampling from the mRNA pool (see Methods). This involves calculating a 'target quality' $R(m|\mu)$ that quantifies the total affinity of each miRNA μ for each fragment m. Specifically, $R(m|\mu)$ corresponds to the enrichment of fragment m among target sites bound by miRNA μ, relative to m's abundance in the mRNA pool. Calculating $R(m|\mu)$ involves summing over all possible hybrid structures between m and μ. The fraction of time fragment m is bound by a RISC loaded with miRNA μ is proportional to the 'target frequency' $R(m|\mu)\pi_\mu$, which additionally depends on the fractions π_μ of RISC complexes loaded with miRNA μ. These fractions, which we call miRNA *priors*, are inferred for each given CLIP data set. The overall probability of immunoprecipitating fragment m relative to its background frequency is then given by $R(m) = \sum_\mu R(m|\mu)\pi_\mu$, and the likelihood of the entire data set by the product $R(D) = \prod_i R(m_i)$ over all observed fragments m_i.

We first tested the procedure on synthetic data sets containing seed-matching sites and 3'-compensatory sites similar to those previously described in the literature [116, 18]. MIRZA successfully inferred the energy parameters that were used in generating these synthetic data sets and perfectly predicted which miRNA was associated with each site (Suppl. Fig. 41). To infer the energy parameters of real miRNA-target interactions from Ago2-CLIP data, we used all miRNAs that were expressed in the HEK293 cells in which the experiments were performed as well as 2988 mRNA regions that were reproducibly crosslinked in at least 3 of 4 Ago2-CLIP data sets from Kishore et al. [95] (see Methods). We extracted 51 nucleotide-long regions centered on the position with the highest number of crosslink-diagnostic mutations (CCRs) and performed 100 parameter optimization runs starting from randomly chosen initial values for all parameters.

Different optimization runs yielded highly reproducible parameter sets (Fig. 18B, Suppl. Fig. 42A). Consistent with the known importance of the seed region, positions 2-7 have the largest positive contribution to the energy (parameters $E_2 \ldots E_7$ in Fig. 18B), followed by positions $13-16$ ($E_{13} \ldots E_{16}$) and $18-19$ (E_{18}, E_{19}). In contrast, hybridization of position 9 (E_9) is strongly disfavoured, as is opening a loop (E_o). Once a loop is opened, symmetric loops (E_{sym}) and bulges in the miRNA (E_μ) are clearly favoured over bulges in the mRNA (E_m).

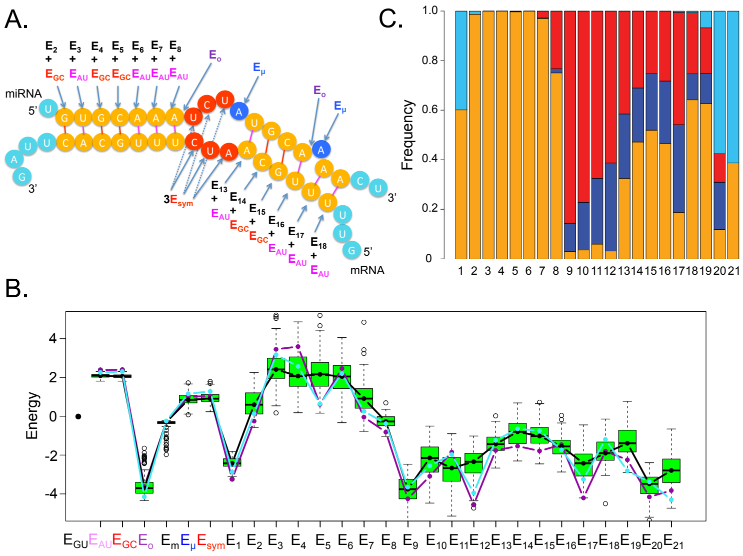


Figure 18: A biophysical model of miRNA-target interaction. **A:** Sketch of miRNA-mRNA hybrid illustrating the way MIRZA assigns a binding energy to the interaction. Nucleotides involved in base-pairing are indicated in orange, symmetric loops in red, bulges in the miRNA in blue and dangling ends in cyan. Arrows point from the independent energy terms to the corresponding structural elements (base pairs, loop openings and extensions, see also Methods). **B:** Summary of energy parameters inferred from 100 independent optimization runs on the Ago2-CLIP data. Green boxes show inter-quartile ranges, 5 and 95 percentiles are indicated by whiskers, *black* dots indicate median values of fitted parameters across the runs. The sets of parameters that yielded the highest and second-highest probability are shown as *purple* and *cyan* dots, respectively. **C:** Summary of the predicted hybrid structures; miRNA positions are labeled on the x-axis and colors indicate the fraction of hybrids in which a given nucleotide is involved in a base-pair (orange), symmetric loop (red), bulge (blue) and dangling end (cyan).

With the fitted parameters we can predict which miRNA μ is most likely to bind each fragment m, as well as the structure of the most likely hybrid between m and μ. Figure 18C statistically summarizes the structures of these predicted hybrids. Strikingly, even though no specific knowledge about miRNA-target interactions went into the inference of its parameters, our model captures several known structural features of miRNA-target interaction such as the predominant binding of the nucleotides in the seed region, the less frequent binding of position 1, and the possibility of compensatory base-pairing at the miRNA's 3′ end. In contrast to general models of RNA-RNA interaction applied to the same data (Suppl. Fig. 42B), MIRZA makes the specific prediction that nucleotides $14-16$ of the miRNA are base-paired with the target roughly 50% of the time and that positions $18-19$ are bound even more than 60% of the time.

Surprisingly, for more than 26% of the most enriched, reproducibly CLIPed sites, the most likely hybrid is non-canonical (Suppl. Fig. 42C). This is especially noteworthy since the more accurate target prediction methods focus solely on canonical sites, and functional non-canonical sites are thought to be rare. However, recent experimental studies hinted that non-canonical sites may be more prevalent, particularly those in which an mRNA nucleotide is bulged out between positions 5 and 6 [26]. Applying MIRZA to the data of Chi et al. [26], we indeed find that, depending on the sample, $9-20\%$ of the predicted miR-124 sites correspond to this particular non-canonical site (Suppl. Table 1). MIRZA however predicts several other types of non-canonical sites, e.g. contiguous pairing of only nucleotides $2-6$, in all CLIP data sets.

MIRZA further infers that the fraction of non-canonical sites is higher for miRNAs with highest abundance in RISC, i.e those with high prior π_μ, and that the fraction of non-canonical sites can be as high as 60% (Fig. 19A). The inferred abundance π_μ correlates significantly with the expression level of the miRNA (Suppl. Fig. 42D), suggesting that the target spectrum of a miRNA depends crucially on its expression level; low expressed miRNAs target mainly high-affinity canonical sites, while highly expressed miRNAs target large numbers of non-canonical sites which, on average, have lower affinity (Suppl. Fig. 42C).

Gene expression analysis shows that the non-canonical sites inferred from the CLIP data are functional, inducing a significant down-regulation of host transcripts upon miRNA transfection (Fig. 19B and E, see Methods). Although sites with higher predicted target quality show stronger down-regulation, even transcripts containing the weakest non-canonical sites show stronger down-regulation compared to transcripts that simply carry seed matches (Fig. 19B). That non-canonical sites show significantly more evolutionary conservation than expected by chance (Suppl. Fig. 43) is further indication of their functionality.

To compare the accuracy of the target sites identified by MIRZA in Ago2-CLIP data with those of miRNA target prediction methods, we analyzed 38 transfection experiments involving 26 different miRNAs [123, 61, 156, 113, 50], comparing the miRNA-induced fold-changes of transcripts predicted by these methods (see Methods and Suppl. Figs. 44, 45, 46 and 47). To assess the ability of a method to identify the most strongly down-regulated targets we sorted its predicted targets by their score, and calculated the median fold-change of the top n targets as a function of n (Fig. 19C). To assess the total number of functional targets predicted by a method we calculated how many more targets were down-regulated compared to the number expected

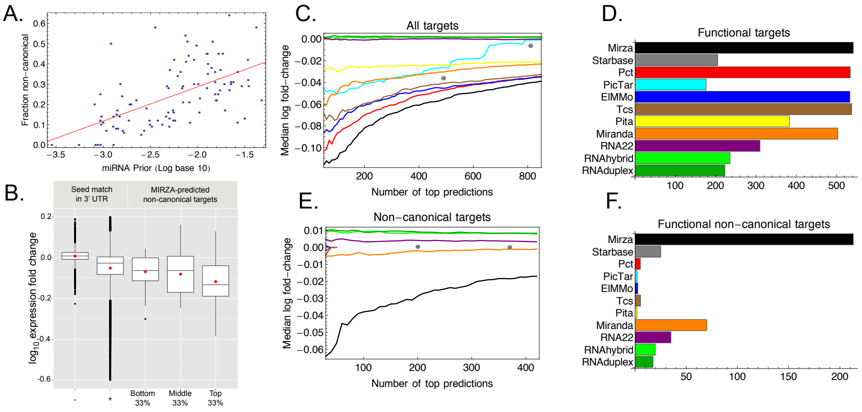


Figure 19: Assessment of the functionality of miRNA targets identified by MIRZA. **A:** Scatter plot showing the correlation between the inferred miRNA prior π_μ (fraction of all silencing complexes loaded with miRNA μ) and the fraction of non-canonical target sites for miRNA μ. The prior is shown on a logarithmic scale. Pearson correlation coefficient $R = 0.58$ (P-value = 2.1×10^{-10}).**B:** Changes in the expression level of mRNAs containing MIRZA-predicted non-canonical binding sites upon transfection of the corresponding miRNAs (expression data from Linsley et al. [123]). Each column shows the distribution of expression changes upon miRNA transfection of a set of transcripts, with the box indicating the interquartile range, the black line the median, the red dot the mean, and the whiskers the 5 and 95 percentiles. The first two columns correspond to transcripts without and with seed matches for the transfected miRNA, and the last three columns correspond to transcripts containing a non-canonical, MIRZA-predicted site with a target quality score among the lowest, middle and highest 33%. **C:** Median log-fold change of targets predicted by the MIRZA (black), TargetScan Pct (red) [44], PicTar (cyan) [63], ElMMo (dark blue) [45], TargetScan context+ (brown) [49], Pita (yellow) [89], Miranda (orange) [15], RNA22 (violet) [135], RNAhybrid (light green), and RNAduplex (dark green) [124], averaged over 38 transfection experiments from 5 studies [123, 61, 156, 113, 50]. The gray dots show fold-changes of targets obtained by intersecting Ago-CLIP sites with computationally-predicted sites (Starbase database, [196]). **D:** Estimated total number of functional targets (see Methods) predicted by the different methods averaged over all transfection experiments. Colors are as in panel C. **E:** Same as panel C, but only considering non-canonical targets, whose 3′ UTRs did not contain a canonical match to the 'seed' region of the transfected miRNAs. **F:** Same as in panel D, but considering only non-canonical targets. Results for individual data sets are shown in Suppl. Fig. 44, and even more detailed results, on individual miRNAs, are shown in Suppl. Figs. 45,46 and 47.

by chance ((Fig. 19D). Although the relative performance of the different methods varies across data sets, MIRZA's predictions show the strongest down-regulation (Fig. 19C) on average, and for the large majority of individual data sets and miRNAs (Suppl. Figs. 44, 45, 46 and 47). Furthermore, MIRZA matches the best methods that use evolutionary conservation (i.e. TargetScan Pct and ElMMo) or the context of the sites (i.e. TargetScan context and Miranda), in the the total number of functional targets that it predicts (Fig. 19D and Suppl. Figs. 44, 45, 46 and 47).

Where MIRZA clearly stands out from other methods is in the prediction of functional non-canonical targets (Fig. 19E, 19F, Suppl. Figs. 44 and 46), whose number is three-fold higher for MIRZA relative to any other method. Furthermore, MIRZA's non-canonical targets undergo a much stronger down-regulation, correlated with their MIRZA score (Fig. 19E), and this performance is consistent across all data sets and individual miRNAs (Suppl. Figs. 44 and 46). The partial overlap between the sites identified for some miRNAs by MIRZA and by algorithms based on conservation or context (Suppl. Fig. 47) suggests that miRNA target prediction could be further improved by combining MIRZA's biophysical model with context and conservation information.

In summary, MIRZA provides a biophysical model of miRNA-target interaction that enables reliable identification not only of canonical but also non-canonical binding sites. MIRZA is made available among the tools provided on our CLIPZ server[1].

5.2 ONLINE METHODS

5.2.1 *Inference of the* MIRZA *model*

We defined a parametrized biophysical model to assign binding free energies to all possible miRNA/mRNA hybrid structures and quantify the binding affinity of different mRNA fragments to the RNA-induced silencing complex (RISC). Because a CLIP experiment does not provide accurate binding frequencies for *all* possible mRNA segments in the transcriptome but rather gives a set of fragments that are enriched relative to the expression of their mRNAs, we extract a set of highly enriched target sites, $m_1, m_2, \ldots, m_n$ of standardized length M from the CLIP data, as described in the Methods. We will make the idealization that the probability of obtaining a particular mRNA fragment m is proportional to the product of the abundance of the mRNA fragment and the fraction of time that the fragment is bound to a RISC. The latter quantity will depend on the binding free energy between the mRNA and RISC. Let $P(m|B)$ denote the "background" abundance of mRNA fragment m in the transcriptome. Let $P(m|IP)$ denote the

[1] http://www.clipz.unibas.ch

probability that when a single bound RISC is immunoprecipitated, this complex will contain a certain mRNA fragment m. This probability depends not only on the relative abundance of m, but also on the relative abundances of the different miRNAs that can interact with the mRNA fragment in RISC. Formally, the probability $P(m|IP)$ can be written as a sum over the probabilities $P(m, \mu|IP)$ that the immunoprecipitated fragment is bound to a RISC containing mature miRNA μ. If we denote by π_μ the fraction of all RISCs that are bound to some target site, that are guided by miRNA μ, then we have

$$P(m|IP) = \sum_{\mu} P(m, \mu|IP) = \sum_{\mu} P(m|\mu)\pi_\mu, \tag{5.1}$$

where $P(m|\mu)$ is the probability that a bound RISC containing miRNA μ is bound to fragment m.

The guide miRNA can form different hybrid structures with an mRNA fragment. Denoting individual hybrid structures by σ and the binding free energy of a RISC-embedded miRNA μ with mRNA fragment m in configuration σ by $E(\sigma, \mu, m)$, from the standard Boltzmann distribution of statistical physics we have that the fraction $P(m|\mu)$ of all RISCs that are loaded with miRNA μ and are bound in configuration σ to mRNA segment m is proportional to $e^{E(\sigma,\mu,m)}P(m|B)$ (note that we set the parameter β of the Boltzmann distribution to 1, for notational simplicity, which can be thought of as setting the scale of the energy parameters). Thus, a RISC complex loaded with miRNA μ is bound to mRNA fragment m with probability

$$P(m|\mu) = \frac{\sum_{\sigma} e^{E(\sigma,\mu,m)}P(m|B)}{\sum_{m',\sigma'} e^{E(\sigma',\mu,m')}P(m'|B)}, \tag{5.2}$$

where the sum in the numerator is over all possible hybrid structures σ, and the sum in the denominator is over all possible hybrid structures and all possible M-nucleotides long mRNA fragments m'. The probability of the entire data is

$$P(D) = \prod_{i=1}^{n} P(m_i|IP), \tag{5.3}$$

where the product is over all n mRNA fragments m_i that are sampled. The probability of observing a fragment m_i when *randomly* selecting fragments from the mRNA pool is just $P(m_i|B)$. Thus, the ratio of probabilities for observing the data under our model as opposed to random sampling is given by

$$R(D) = \prod_{i=1}^{n} \frac{P(m_i|IP)}{P(m_i|B)} = \prod_{i=1}^{n} R(m_i). \tag{5.4}$$

The ratios $R(m_i)$ quantify to what extent the observation of m_i is explained by miRNA binding, i.e. they give the *enrichment* of fragment

m_i when immunoprecipitating with RISC relative to its abundance in the mRNA pool.

Finally, we will make use of the following *partition function* notation

$$Z(\mu) = \sum_{m,\sigma} e^{E(\sigma,\mu,m)} P(m|B) \tag{5.5}$$

and

$$Z(m,\mu) = \sum_{\sigma} e^{E(\sigma,\mu,m)}. \tag{5.6}$$

With this notation we have

$$R(m) = \frac{P(m|IP)}{P(m|B)} = \sum_{\mu} \frac{P(m|\mu)}{P(m|B)} = \sum_{\mu} R(m|\mu)\pi_{\mu} = \sum_{\mu} \frac{Z(m,\mu)}{Z(\mu)}\pi_{\mu}. \tag{5.7}$$

5.2.2 *Target quality and target frequency*

The quantity $R(m|\mu)$ represents the ratio of the probability that a RISC guided by miRNA μ binds to segment m, and the background probability of isolating segment m, $P(m|B)$. In other words, $R(m|\mu)$ is the enrichment of fragment m among all fragments bound to a RISC loaded with miRNA μ relative to its background frequency $P(m|B)$. Because $R(m|\mu)$ quantifies the quality of segment m for miRNA μ (i.e. relative to all other possible target segments) we will refer to it as the *target quality*. Note, however, that for a given segment m, the miRNA with the highest target quality $R(m|\mu)$ is not necessarily the miRNA that most frequently associates with segment m because this latter quantity depends also on the relative abundances π_μ of RISCs that are loaded with different miRNAs. As can be seen from equation (5.7), the fraction of time that segment m is bound by miRNA μ and, consequently, the miRNA that most frequently binds to segment m, is the one that maximizes the product $R(m|\mu)\pi_\mu$. We will refer to $R(m|\mu)\pi_\mu$ as the *target frequency* of miRNA μ for segment m.

5.2.3 *Parameterization of the binding energies*

Ignoring the possibility that the miRNA or the mRNA fragment form internal structures (base-pairing within themselves), our model assumes that each possible hybrid structure σ consists of one or more hybridized pairs of nucleotides that are separated by unpaired nucleotides, forming either symmetrical or asymmetrical loops, depending on whether the number of unpaired nucleotides in the miRNA and mRNA are the same or different. A hybrid σ can then be uniquely represented using the following set of 'moves':

1. an initial hybridized pair (i, j), i.e. position i in the miRNA hybridized to position j in the mRNA fragment,
2. addition of another hybridized pair immediately following the current pair,
3. opening a loop,
4. adding a symmetric pair of unhybridized nucleotides to the loop,
5. adding an unpaired nucleotide in the mRNA fragment,
6. adding an unpaired nucleotide in the miRNA.

To ensure that each possible hybrid can only be realized in one way using these moves, we make the convention that asymmetric additions to loops can only be followed by more asymmetric additions of the same type, or by a hybridized pair. Similarly, symmetric additions can only be followed by additional symmetric additions, by an asymmetric addition, or a hybridized pair. Hybrids have to end in a hybridized pair, and the remaining nucleotides in mRNA fragment and miRNA are considered "dangling ends".

For each possible hybrid that can be constructed as described above, we assume that the binding energy can be decomposed into a *structural* and a *sequence* component:

$$E(\sigma, m, \mu) = E_{\text{struc}}(\sigma) + E_h(\sigma, m, \mu). \tag{5.8}$$

The structural contributions to the energy are determined from the 'moves' and are an energy E_o for every loop that is opened, an energy E_{sym} for symmetrically extending a loop by 1 base in the miRNA and 1 base in the mRNA, an energy E_μ for asymmetrically extending a loop by an unpaired base in the miRNA, an energy E_m for asymmetrically extending a loop by an unpaired base in the mRNA fragment, and an energy E_i when position i in the miRNA is hybridized. The latter reflects the constraints that the Argonaute protein imposes on the embedded miRNA, e.g. through the accessibility of the corresponding position of the miRNA when it is inside RISC. Without loss of generality, dangling bases in mRNA and miRNA per definition are assigned an energy $E_d = 0$. Thus, the structural part $E_{\text{struc}}(\sigma)$ depends on the number of loops, their sizes, their (a)symmetry, and on the positions in the miRNA that are hybridized. This dependency on miRNA position enters through the energies E_i of the hybridized positions.

The sequence-dependent part of the energy consists of a sum of energy contributions for each hybridized pair, with $E_{\alpha\beta}$ being the energy contribution for hybridizing nucleotide α in the mRNA to nucleotide β in the miRNA. If we denote by h the set of miRNA positions that are hybridized in structure σ, we have

$$E_h(\sigma, m, \mu) = \sum_{i \in h} E_{m_i \mu_i}, \tag{5.9}$$

with m_i the nucleotide occurring at the position in the mRNA segment hybridized to miRNA position i and μ_i the nucleotide at position i of the miRNA. Although in the most general case we would need 16 parameters to describe these contributions, we have only considered the usual base-pairing interactions A-U/U-A, C-G/G-C and G-U/U-G, which we described by parameters $E_{AU} = E_{UA}$, $E_{CG} = E_{GC}$, and $E_{UG} = E_{GU}$. We assign all other combinations a very negative energy, i.e. $-\infty$, such that they have zero probability of occurrence.

5.2.4 *Redundancies*

To infer the energy parameters the observed data D, it is important to determine whether our parameterization contains redundancies, i.e. if there are global transformations of the parameters that would leave the overall likelihood ratio $R(D)$ invariant. In the model described above, a redundancy results from the fact that, for every hybridized base pair (α, β) there is a sequence-dependent contribution $E_{\alpha\beta}$ and a structural contribution E_i from the hybridized position i in the miRNA. Thus, if we replace

$$E_{\alpha\beta} \to E_{\alpha\beta} + c, \tag{5.10}$$

for all pairs (α, β), and at the same time replace

$$E_i \to E_i - c, \tag{5.11}$$

then all energies $E(\sigma, m, \mu)$ remain unchanged. To remove this redundancy, we assign one of these parameters a "neutral" value. We chose to set $E_{GU} = 0$. The energies E_d of the dangling ends are set to zero as well to avoid redundancies in the parameterization.

As detailed below, we will fit all the energy parameters of the model by optimizing the likelihood of the observed CLIP data. The reader may wonder why certain parameters, such as the energies associated with base pairing, are not simply set to experimentally estimated values such as those that are used in RNA secondary structure prediction algorithms. It is important to stress that the energy parameters that we are inferring here are the effective contributions of various structural components (e.g. base pairs, loops) *in the context of the RISC complex*. That is, the interaction of the miRNA and mRNA target will be likely be strongly influenced by the context provided by this protein complex, and it is therefore not *a priori* clear what the contributions of different base pairs and loops should be.

5.2.5 *Partition function*

The partition function $Z(\mu) = \sum_{\sigma,m} e^{E(\sigma,m,\mu)} P(m|B)$ can be derived in terms of the above-defined parameters. Considering separately the structure and sequence components of the energy (see equation 5.8)

$$E(\sigma, m, \mu) = E_{\text{struc}}(\sigma) + E_h(\sigma, m, \mu),$$

we can write

$$Z(\mu) = \sum_{\sigma} \left[e^{E_{\text{struc}}(\sigma)} \sum_{m} P(m|B) e^{E_h(\sigma,m,\mu)} \right]. \tag{5.12}$$

In order to be able to recursively calculate this partition function we approximate the distribution $P(m|B)$ using a simple model that takes only the overall nucleotide frequencies into account. That is, we assume that $P(m|B) = \prod_{j=1}^{M} w(m_j)$, where m_j is the nucleotide that occurs at position j in the mRNA fragment and $w(\alpha)$ is the frequency of nucleotide α in the entire set of 3'UTRs. We use the base frequencies within 3' UTRs rather than entire mRNAs since miRNAs are known to preferentially bind to 3' UTRs. The sequence-dependent contributions can then be separated into a product over the non-hybridized and the hybridized positions yielding

$$Z(\mu) = \sum_{\sigma} \left[e^{E_{\text{struc}}(\sigma)} \sum_{m} \left(\prod_{i \notin h} w(m_i) \right) \left(\prod_{i \in h} w(m_i) e^{E_{m_i \mu_i}} \right) \right], \tag{5.13}$$

where h is the set of positions in the mRNA that are hybridized, and μ_i is the nucleotide in the miRNA at position i. Noting that the order of the sum and the products can be exchanged and that, for all non-hybridized positions, the sum over the nucleotide in the mRNA gives a factor 1 at each position. We then have

$$Z(\mu) = \sum_{\sigma} \left[e^{E_{\text{struc}}(\sigma)} \prod_{i \in h} \left(\sum_{\alpha} w(\alpha) e^{E_{\alpha \mu_i}} \right) \right]. \tag{5.14}$$

The expression

$$\sum_{\alpha} w(\alpha) e^{E_{\alpha \mu_i}}, \tag{5.15}$$

is the average statistical weight associated with hybridized base μ_i in the miRNA, averaged over the probabilities $w(\alpha)$ that a letter α occurs in the mRNA. For ease of notation below, we denote this quantity by

$$e^{E_{\mu_i}} = \sum_{\alpha} w(\alpha) e^{E_{\alpha \mu_i}}. \tag{5.16}$$

The partition function finally takes the form

$$Z(\mu) = \sum_{\sigma} e^{E_{\text{struc}}(\sigma) + \sum_{i \in h} E_{\mu_i}}. \tag{5.17}$$

5.2.6 *Recursion formulas*

There are two partition sums that we have to determine recursively: $Z(\mu)$, whose form we just derived and $Z(m,\mu)$, that takes the form

$$Z(m,\mu) = \sum_{\sigma} e^{E(\sigma,\mu,m)} = \sum_{\sigma} \left[e^{E_{\text{struc}}(\sigma)} \prod_{i \in h} e^{E_{m_i \mu_i}} \right], \tag{5.18}$$

for each mRNA fragment m and each possible binding miRNA μ. We can introduce a *weight vector*

$$w(m_i|\mu_i) = \frac{w(m_i)e^{E_{m_i\mu_i}}}{\sum_{\alpha} w(\alpha)e^{E_{\alpha\mu_i}}}, \tag{5.19}$$

which assigns probabilities to different letters m_i in the mRNA given a miRNA letter μ_i. Together with the definition (5.16) of E_{μ_i} we can then write

$$Z(m,\mu) = \sum_{\sigma} \left[e^{E_{\text{struc}}(\sigma)} \prod_{i \in h} \frac{w(m_i|\mu_i)}{w(m_i)} e^{E_{\mu_i}} \right]. \tag{5.20}$$

This form nicely emphasizes that the dependence on the mRNA sequence m comes in through the ratios of probabilities $w(m_i|\mu_i)/w(m_i)$ of observing mRNA fragment base m_i given that it is hybridized to miRNA base μ_i and the probability $w(m_i)$ under our *background* model.

In section 5.2.3, we introduced a set of 'moves' to generate all possible hybrid structures σ. These moves can be used to recursively calculate the partition sums $Z(\mu)$ and $Z(m,\mu)$. Let $F_{\alpha}(i,j)$ denote the partition sum of all possible sub-structures of the first i bases in the miRNA and first j bases in the mRNA fragment, that end with move α, where α can be either $\alpha = h$ when adding a hybridized pair $\alpha = \text{sym}$ when adding a pair of symmetrically looped out nucleotides (i.e. one in the miRNA and one in the mRNA fragment), $\alpha = m$ when adding an unpaired nucleotide in the mRNA fragment, and $\alpha = \mu$ when adding an unpaired nucleotide in the miRNA. The sums $F_{\alpha}(i,j)$ can be determined using the following recursion relations.

$$\begin{aligned} F_h(i,j) = e^{E_i + E_{\mu_i}}[1 + F_h(i-1,j-1) \\ + F_{\text{sym}}(i-1,j-1) + F_m(i-1,j-1) + F_{\mu}(i-1,j-1)], \end{aligned} \tag{5.21}$$

where the first term, 1, corresponds to the case in which the pair (i,j) is the first hybridized pair. The other terms correspond, respectively, to extending from a previously hybridized pair, following a pair of symmetrically looped out nucleotides, following a 'bulged out' nucleotide in the mRNA fragment, and following a bulged out nucleotide in the miRNA.

When ending with a symmetric extension we have

$$F_{\text{sym}}(i,j) = e^{E_{\text{sym}}} \left[e^{E_o} F_h(i-1,j-1) + F_{\text{sym}}(i-1,j-1) \right], \tag{5.22}$$

where the two terms correspond to opening a new loop or extending from a previous pair of symmetrically looped out nucleotides. Note that, by construction, symmetric extensions are not allowed after asymmetric extensions. To end with an unpaired nucleotide in the mRNA fragment we have

$$F_m(i,j) = e^{E^m_{as}} \left[e^{E_o} F_h(i,j-1) + F_{sym}(i,j-1) + F_m(i,j-1) \right]. \quad (5.23)$$

Finally, to end with an asymmetric extension of the miRNA we have

$$F_\mu(i,j) = e^{E^\mu_{as}} \left[e^{E_o} F_h(i-1,j) + F_{sym}(i-1,j) + F_\mu(i-1,j) \right]. \quad (5.24)$$

The *boundary conditions* of these recursion relations are that $F_x(i,j) = 0$ whenever $i < 0$ or $j < 0$, for all values of x, and where $i = 0$ and $j = 0$ correspond to the first positions in mRNA and miRNA.

The full partition sum is now given by

$$Z(\mu) = \sum_{i=0}^{L_\mu - 1} \sum_{j=0}^{L_m - 1} F_h(i,j), \quad (5.25)$$

which corresponds to summing over all structures that end with a base-pair at (i,j), the remaining nucleotides in the miRNA and the mRNA being dangling ends that do not contribute to the free energy of interaction.

For the partition sum $Z(m,\mu)$ we have very similar recursion relations. We let $H_\alpha(i,j)$ denote the partition sum over all possible sub-structures of the first i bases in the miRNA and first j bases in the mRNA fragment that end in *state* α. The recursion relations are in fact exactly the same as those for the quantities $F_\alpha(i,j)$, except for when $\alpha = h$, i.e. when ending in an hybridized pair of nucleotides (i,j). For this case we have the relation

$$H_h(i,j) = \frac{w(m_j|\mu_i)}{w(m_j)} e^{E_i + E_{\mu_i}} [1 + H_h(i-1,j-1)$$
$$+ H_{sym}(i-1,j-1) + H_m(i-1,j-1) + H_\mu(i-1,j-1)]. \quad (5.26)$$

The final partition function is again calculated as:

$$Z(m,\mu) = \sum_{i=0}^{L_\mu - 1} \sum_{j=0}^{L_m - 1} H_h(i,j). \quad (5.27)$$

5.2.7 *Definition of best hybrids*

To gain insight into the interactions that our model predicts we determined, for each (miRNA, target site) pair (μ and m) the hybrid structure σ that maximizes the energy $E(\sigma,\mu,m)$ which corresponds to the most likely hybrid structure formed by miRNA μ with target site m. We used for this purpose recursion relations analogous to those for calculating the partition sums above, with summation being replaced by maximization. That is, let $B_\alpha(i,j)$ denote the energy of the best hybrid structure for the subsequences up to nucleotides i and j in miRNA and mRNA fragment, ending in state α. We then have the recursion relations

$$B_h(i,j) = \frac{w(m_j|\mu_i)}{w(m_j)} e^{E_i + E_{\mu_i}} \max\left[1, B_h(i-1,j-1), B_{sym}(i-1,j-1), B_m(i-1,j-1), B_\mu(i-1,j-1)\right] \tag{5.28}$$

$$B_{sym}(i,j) = e^{E_{sym}} \max\left[e^{E_o} B_h(i-1,j-1), B_{sym}(i-1,j-1)\right], \tag{5.29}$$

$$B_m(i,j) = e^{E_{as}^m} \max\left[e^{E_o} B_h(i,j-1), B_{sym}(i,j-1), B_m(i,j-1)\right], \tag{5.30}$$

and

$$B_\mu(i,j) = e^{E_{as}^\mu} \max\left[e^{E_o} B_h(i-1,j), B_{sym}(i-1,j), B_\mu(i-1,j)\right]. \tag{5.31}$$

The *best hybrid* is then constructed by tracing back the 'moves' that lead to the hybrid with the maximum energy.

5.2.8 *Fitting the fraction of RISC complexes carrying specific miRNAs*

Given a fixed set of energy parameters, we use the recursion relations to determine the partition sums $Z(\mu)$ and $Z(m,\mu)$. The final likelihood-ratio $R(D)$ also depends on the fractions π_μ of bound RISCs that are loaded with miRNA μ. Given the $Z(\mu)$ and $Z(m,\mu)$, it is relatively straightforward to determine the fractions π_μ that maximize the likelihood-ratio $R(D)$.

The derivative of the $\log$-likelihood ratio log[R(D)] with respect to one of the fractions is given by

$$\frac{\partial \log[R(D)]}{\partial \pi_\mu} = \sum_i \frac{R(m_i|\mu)}{R(m_i)}. \tag{5.32}$$

The maximum of $R(D)$ under the constraint that

$$\sum_\mu \pi_\mu = 1, \tag{5.33}$$

thus satisfies

$$\pi_\mu = C \sum_i \frac{R(m_i|\mu)\pi_\mu}{R(m_i)}, \tag{5.34}$$

where C is a normalizing constant. We use this expression to determine the optimal fractions π_μ using 'expectation maximization', see e.g. [32]. That is, given a current set of fractions π_μ, we determine for each μ

$$Q(\mu) = \sum_i \frac{R(m_i|\mu)\pi_\mu}{R(m_i)}. \tag{5.35}$$

We then set a new set of fractions

$$\pi_\mu = \frac{Q(\mu)}{\sum_{\mu'} Q(\mu')} \tag{5.36}$$

and iterate until the convergence of π_μ. Because $\log[R(D)]$ is a convex function of the π_μ, the expectation-maximization is guaranteed to find the global optimum.

5.2.9 *Implementation of the parameter optimization*

We implemented our MIRZA algorithm in C++, in an object-oriented framework. It takes as input fasta-formatted files of mRNA fragments and miRNA sequences. To avoid biases introduced by the slight differences in length of different miRNAs we trimmed all miRNA sequences to 21 nucleotides. We optimized the parameters of our biophysical model through simulated annealing for which we used the GNU scientific library [2] and an object-oriented library for numerical programming in C++ (O2SCL [3]). For efficiency, we further used the Open Multi-Processing Architecture (OpenMP [4]) which supports multi-platform shared-memory parallel programming in C/C++ and Fortran. The parameters that we optimized were:

- the base-pairing energies E_{AU} and E_{CG},
- the loop energies E_o, E_{sym}, E^{μ}_{as}, E_m and
- the positional hybridization energies, E_i where $i = 1, \ldots, 21$.

For both the synthetic and Ago2 CLIP data-sets we performed multiple simulated annealing runs starting from various initial conditions and analyzed the reproducibility of the fitted parameters (see main text and supplementary material).

5.2.10 *Argonaute 2 CLIP experimental data sets*

Of the recently reported data sets of Argonaute 2 binding sites, those generated with PAR-CLIP (**P**hoto**a**ctivatable-**R**ibonucleoside-Enhanced **C**ross**l**inking and **I**mmuno**p**recipitation) exhibit frequent diagnostic mutations (transition of uridine to cytidine), typically in the center of the miRNA-target site hybrid [67, 95].

[2] http://www.gnu.org/s/gsl/

[3] http://o2scl.sourceforge.net/

[4] http://openmp.org/wp/

Within each single-linkage cluster that contained sites from at least 3 of 4 Ago2 CLIP/PAR-CLIP samples from Kishore et al. [95] that were generated with various protocols (GEO database accessions GSM714642, GSM714644, GSM714645, GSM714647) we identified the nucleotide with the highest frequency of crosslink-diagnostic mutations, and extracted regions of 51 nucleotides centered on the position of crosslink (crosslink-centered regions, CCRs). A set of 2988 'high-confidence' CCRs that were both among the top 3000 in terms of coverage by sequence reads and in terms of enrichment in the read coverage relative to the read coverage of the same region in HEK293 mRNA-Seq samples were retained for further analyses (Supplementary Table 2 [5]).

5.2.11 *miRNA transfection data for functional analysis of predicted sites*

To investigate the functionality of canonical and non-canonical targets predicted by various methods we used published microarray datasets of changes in gene expression following the transfection of different miRNAs. We selected data sets corresponding mostly to miRNAs that are expressed in HEK293 cells from which CLIP data has been obtained. We further retained data from 'successful' transfection experiments, meaning those in which the mRNAs carrying canonical sites for the transfected miRNA in their 3′ UTR were significantly down-regulated compared the remaining other mRNAs (Wilcoxon's rank-sum test on $\log_2$ fold changes, p-value cut-off of 0.001) and discarded the other data sets. The 5 data sets that we thus used are summarized below.

- Linsley et al. [123] transfected 11 miRNAs (miR-16, miR-15a, miR-106b, miR-20a, miR-103, miR-17, miR-20a, and let-7c) in HCT116 and DLD-1 cell lines, each in duplicate. The processed differential expression data from the GEO database (accession GSE6838, experiments GSM156557, GSM156558, GSM156580, GSM156544, GSM156543, GSM156576, GSM156545, GSM156549, GSM156546, GSM156550, GSM156553, GSM156555, GSM156532, GSM156541, GSM156554 and GSM156556) together with the probe to transcript mapping provided by the authors as a SOFT formatted file were downloaded. Probes associated to Refseq transcripts according to the annotation were kept for subsequent analysis. Differential expression at the gene-level was obtained by mapping RefSeq IDs to Entrez Gene IDs using the RefSeq database downloaded on January 11th, 2007. For each gene, fold-changes were averaged over the duplicate experiments.

- Grimson et al. [61] transfected 9 miRNAs (miR-122, miR-128, miR-132, miR-133a, miR-142-3p, miR-148b, miR-181a, miR-7, and miR-9) in HeLa cells, profiling mRNA expression 12h and

[5] http://www.nature.com/nmeth/journal/v10/n3/extref/nmeth.2341-S2.xlsx

24h post-transfection. We retrieved the processed differential expression data from GEO (GSE8501) and then applied the same analysis as for the data of [123], on the fold-change data from the 24h time point.

- Leivonen et al. [113] transfected miR-18a, miR-193b, miR-302c and miR-206 into MCF7 cells. We again retrieved the processed data from the GEO database and computed average expression levels per Entrez Gene ID. We then computed the $\log_2$ fold change in expression levels upon miRNA transfection as compared to scrambled pre-miR control.
- Selbach et al. (2008)[156] transfected 5 miRNAs in HeLa cells (miR-155, miR-16, miR-1, miR-30a, and let-7b). The CEL files of Selbach et al. (2008) were downloaded from http://psilac.mdc-berlin.de/download/. Of these five miRNA transfections in HeLa, we excluded the let-7b experiment because of the reported negative feed-back of let-7b on the RNAi pathway due to direct targeting of Dicer [175, 156, 72].
- Finally, we obtained the CEL files of the miR-26b and miR-98 overexpression in HeLa cells performed by Gennarino et al. [50] from GEO (accession GSE12100).

We imported the CEL files into the R software [6] using the BioConductor affy package [51]. The probe intensities were corrected for optical noise, adjusted for non-specific binding, and quantile normalized with the gcRMA algorithm [195]. Probe sets with more than two probes mapping ambiguously (more than one 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 $\log_2$ fold changes to obtain an expression change per gene.

Altogether these 5 data sets cover changes in gene expression in 38 different transfection experiments involving 26 distinct miRNAs.

5.2.12 *Comparison of miRNA target prediction methods*

Some methods predict miRNA target *sites*, while others predict *transcripts* that are targeted by individual miRNAs. To be able to compare these heterogeneous predictions, we worked at the level of transcripts, and for methods that predicted target *sites* we assumed that the transcript score is given by the highest score of any predicted site in that transcript. The methods that we considered were:

- ElMMo [46] (http://www.mirz.unibas.ch/ElMMo3/), which estimates the selection pressure on individual sites through comparative genomics.

[6] http://www.R-project.org

- PicTar [63] (http://dorina.mdc-berlin.de/rbp_browser/hg18.html), another comparative genomics-based method whose predictions are widely used.

- TargetScan P_{ct} [44] (http://www.targetscan.org), also a method that uses evolutionary conservation. Based on previous evaluations this method is considered one of the most accurate methods for identifying functional target sites.

- TargetScan context+ [61, 49] (http://www.targetscan.org), which predicts miRNA target sites based on the sequence context in which they occur in their host transcripts.

- MIRANDA [15] (http://www.microrna.org/microrna/getDownloads.do). The current version of this method, *mirSVR*, uses support vector regression based on a list of features of both the miRNA and its putative target. MIRANDA provides 4 separate files of targets, depending on whether targets are filtered based on mirSVR score and/or conservation. We used the 'S' sets of targets that are filtered by score, irrespective of conservation.

- PITA, release of 31-Aug-08 [89], which computes an energy of interaction between miRNA and target site taking into account the structural accessibility of the target site; we extracted predictions for the miRNAs of interest from the web site, http://genie.weizmann.ac.il/pubs/mir07/mir07_dyn_data.html, queried with default parameters.

- RNAduplex, which computes the minimum free energy of hybridization between the two RNA strands [124]. We downloaded RNAduplex as part of the Vienna RNA package from http://www.tbi.univie.ac.at/RNA/RNAduplex.html and applied it to the entire set of representative 3′ UTRs of human genes.

- RNAhybrid [148], which uses an approach similar to RNAduplex. We downloaded RNAhybrid from the server hosted by the University of Bielefeld (http://bibiserv.techfak.uni-bielefeld.de) and applied it to the entire set of representative 3′ UTRs of human genes.

- RNA22 [135], a method based on statistical over-representation of miRNA-complementary motifs. Current genome-wide predictions of this method were obtained from the authors.

We further included lists of targets of miRNAs from the Starbase database [196] (http://starbase.sysu.edu.cn/), which intersects Ago2 CLIP sites with miRNA target predictions by TargetScan, PicTar, Miranda, PITA, and RNA22. Starbase does not provide a default sorting of predicted sites but allows users to manipulate stringency parameters. We downloaded target lists using the default settings of the database, and using the most inclusive settings which maximizes the total number of predicted sites.

Since different methods may use different transcript collections as a basis for their predictions, we decided to compare predictions at the level of Entrez genes. For ElMMo, PicTar, TargetScan, MIRANDA, PITA, and RNA22 we collected, for each Entrez gene, all transcripts associated with the gene and defined the target score as the highest score among all transcripts in the set.

For RNAhybrid and RNAduplex, we predicted miRNA-complementary sites transcriptome wide. For this purpose, we selected a representative 3′ UTR for each Entrez Gene that had a Refseq transcript in the January 18, 2011 release of the Refseq database. We chose as representative 3′ UTR the 3′ UTR of the longest transcript among those that were represented in Refseq, had an annotated 5′ UTR, 3′ UTR, and CDS, and were associated with the corresponding gene. We scanned each 3′ UTR with windows of 50 nucleotides, shifting by 25 nucleotides at a time, and predicted the minimum free energy of interaction between the miRNA and each window. We then defined the 'transcript score' as the minimum free energy over all windows from the 3′ UTR of a give transcript. For MIRZA, the target score of a transcript from the representative set was defined as the sum of the logarithms of the target qualities of all sites occurring in the transcript.

5.2.13 *Median fold-changes*

To test the accuracy of the target predictions of each method we used the data sets of miRNA transfection experiments as described in section 5.2.11. For each transfection experiment, and each method, we sorted all predicted target genes by score and filtered out all genes for which no fold-change data was available in the corresponding data-set. We then determined, as a function of the number n of top predicted targets, the median $\log$ fold-change $\text{lm}(n)$ of these targets in response to miRNA transfection. Lower median fold-changes thus indicate that a method predicts targets that are more strongly down-regulated upon transfection of the miRNA. For each of the 5 data-sets, we calculated average median $\log$ fold-changes $\langle \text{lm}(n) \rangle$ by averaging the functions $\text{lm}(n)$ of each of the transfection experiments in individual data sets. We also calculated an average over all 38 transfection experiments.

5.2.14 *Number of functional targets*

Besides calculating median $\log$ fold-changes we also determined, for each miRNA and each method, the *fraction* $f(n)$ of the top n predicted targets that were down-regulated, as a function of the number n of top predicted targets. We used the functions $f(n)$ to estimate the total number of functional targets as follows. For each data set, we first determined the total fraction f_{tot} of down-regulated transcripts among all transcripts for which fold-change data was measured. Typically, f_{tot} is close to 50%. Thus, if we were to make random predictions, we expect a fraction f_{tot} of the predicted targets to be down-regulated. If $f(n)$ is considerably larger than f_{tot}, this indicates that there must be true targets among the n predicted targets. Notice that, if a fraction $\rho(n)$ of the n predicted targets are 'true targets', and using that fact that true targets must be down-regulated per definition, then the total fraction $f(n)$ of down-regulated targets will be

$$f(n) = \rho(n) + f_{\text{tot}}\left(1 - \rho(n)\right). \tag{5.37}$$

From this we can estimate the total number of functional targets as

$$n_{\text{func}}(n) = n\rho(n) = n\frac{f(n) - f_{\text{tot}}}{1 - f_{\text{tot}}}. \tag{5.38}$$

For each method and each transfection experiment, we determine the total number of functional targets by maximizing $n_{\text{func}}(n)$ over n, i.e. we choose the number n of top predicted targets such that $n_{\text{func}}(n)$ is maximal.

$$n_{\text{func}} = \max_{n}\left[n\frac{f(n) - f_{\text{tot}}}{1 - f_{\text{tot}}}\right]. \tag{5.39}$$

For each method, we also determined the average number of functional targets $\langle n_{\text{func}}\rangle$ for each of the 5 data-sets, by averaging n_{func} over the transfection experiments in a data-set. We also calculate an overall $\langle n_{\text{func}}\rangle$ averaging over all 38 transfection experiments. Finally, all these calculations were also performed restricting the targets to those transcripts that do not contain a canonical match to the seed sequence of the miRNA, as described in the next section.

5.2.15 *Non-canonical binding sites*

To identify non-canonical target sites among the CLIPed sites, we used the following stringent procedure. We first predicted with MIRZA the miRNA μ with which each mRNA fragment m most likely interacted, i.e. the miRNA for which the mRNA fragment had the highest target frequency $R(m|\mu)\pi_{\mu}$. Next, we determined the optimal hybrid σ for this miRNA-mRNA target pair with the recursion relations described above, and based on these hybrids we divided the set of mRNA fragments into 2 subsets:

- canonical sites [10], which base-paired contiguously with nucleotides $2-8$ of the miRNA **OR** had an exact match to positions $2-7$ of the miRNA followed by an adenine (which would be positioned opposite position 1 of the miRNA).

- non-canonical sites, for which the above condition was not satisfied.

We then identified transcripts that contained a single, non-canonical CLIPed site for the transfected miRNA and retained those transcripts that did not additionally contain a canonical *seed match* (as defined above) anywhere in the 3′UTR. We used in this search the 3′ UTRs of representative transcripts from Kishore et al. [95]. This procedure gave us a very conservative set of transcripts on which the miRNA was very likely to act on a non-canonical site. We sorted the non-canonical sites based on their *target quality* $R(m|\mu)$ with respect to the transfected miRNA μ and then divided the set into 3 subsets of equal size, corresponding to the top 33%, the middle 33% and the bottom 33% in terms of the target quality. To resolve issues of differences between genome and transcriptome annotations, we investigated the change in expression at the level of genes. That is, we mapped transcripts to corresponding genes in the Entrez database of NCBI. Finally, we compared the expression level changes between genes containing sites within each subset and genes whose representative transcripts did not contain a seed match in the 3′UTR or did contain a seed match (irrespective of whether it was CLIPed) in the 3′UTR. For each gene, we computed the average $\log$-fold change across replicate transfection experiments.

For the comparison of prediction accuracy of the different target prediction methods we defined non-canonical targets as follows. For each miRNA, we scanned all 3′ UTRs of RefSeq transcripts associated with each Entrez gene for a canonical match to the miRNA. All Entrez genes for which such a seed match was detected are considered canonical targets by default, i.e. irrespective of where in the 3′ UTR the various methods predicted sites or which of the RefSeq transcripts contained such a site. Thus, non-canonical target genes of a given miRNA are those for which the 3′ UTRs of associated RefSeq transcripts do not contain a canonical match to the seed sequence.

5.2.16 *Representation of non-canonical binding modes among CLIP sites*

To determine the prevalence of specific non-canonical binding 'modes' in CLIP data sets, we extracted sites as follows. From each of the 4 Ago2 data sets from Kishore et al. [95], and from the three mouse brain Ago2 HITS-CLIP data sets (libraries prepared from the 130 kDa band) of Chi et al. [23] we extracted the 5000 sites with the highest coverage by reads. We also extracted the 5000 most enriched sites, relative to the expression of the corresponding mRNAs in an mRNA-seq sample that we prepared from HeLa cells, from the two samples from Chi et al. [23] that were obtained after miR-124 transfection in HeLa cells. We applied the MIRZA model to each of these sets of putative Ago2 binding sites to determine the miRNAs that most likely guided the interaction with the site and the hybrid with the highest score and we used this to determine the relative proportions of individual 'binding modes' (e.g. that with a bulge at the 'pivot position' [26]) among these hybrids.

5.3 AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: Erik van Nimwegen and Mihaela Zavolan. Performed the experiments: Mohsen Khorshid, Jean Hausser. Analyzed the data: Jean Hausser, Mohsen Khorshid, Erik van Nimwegen, Mihaela Zavolan. Wrote the paper: Jean Hausser, Mohsen Khorshid, Mihaela Zavolan and Erik van Nimwegen.

5.4 ACKNOWLEDGMENTS

We are grateful to Niko Beerenwinkel and Sven Bergmann for comments in the initial stages of this work. We are also thankful to Andreas R. Gruber and the other members of the Zavolan group for providing input and feedback on the algorithm and the manuscript and to Alessandro Crippa for help with the code distribution. Mohsen Khorshid was supported by Swiss National Science Foundation ProDoc grant PDFMP3_123123 to Mihaela Zavolan and Erik van Nimwegen.

Part IV

CONCLUSION AND FUTURE DIRECTIONS

6 CONCLUSION AND FUTURE DIRECTIONS

ABSTRACT

In the following sections, the main findings of the thesis are discussed and an outlook of where future work could be taken up, is given.

6.1 SUMMARY OF THE RESULTS

PAR – CLIP, *a powerful cell-based crosslinking approach to determine at high resolution and transcriptome-wide the binding sites of cellular RBPs and miRNPs*

We showed that application of photoactivatable nucleoside analogs to living cells facilitates RNA-protein crosslinking and transcriptome-wide identification of RBP and RNP binding sites. As a proof of concept, PAR-CLIP was successfully applied on RNA binding proteins such as Pumilio-2, supporting evidence from previous studies. Furthermore, the method was applied to various other RNA binding proteins (such as Argonaute family, IGF2BP, HuR) for which the binding specificity was not fully characterized and we were able to determine and validate their consensus sequence binding motif for these RBPs.

Novel approaches to identify binding sites from CLIP data

CLIP experiments are always contaminated with non-crosslinked RNAs (e.g. as shown by consistent identification of rRNAs, tRNAs, and microRNAs). In fact, they have a certain amount of isolated RNAs which do not represent regulatory binding sites. We showed that the crosslinked nucleotides induce a specific mutational signature in the sequenced binding sites relative to the reference genome. This mutational signature [1] can be used to separate the crosslinked binding sites from other RNA fragments that are captured during the experiment thereby allowing accurate identification of RBP binding sites.

We provided a quantitative analysis to improve the quality and reproducibility of the CLIP methods. For example, using extensive digestion with sequence-specific ribonucleases (e.g. T1) could dramatically affect the obtained binding sites. Furthermore, we proposed using milder nuclease digestion conditions as a solution to the problem that could reduced this effect. We also suggested a set of novel bioinformatics methods to improve quality of data analysis of the obtained RNAs.

CLIPZ, *a database and analysis environment for experimentally-determined binding sites of RNA-binding proteins*

It is not surprising any more to obtain very large amounts of data in a short time as results of biological experiments and in fact this capability has changed the field of biology. Like many high-throughput

[1] specifically mutations of the crosslinked *uridines* to *cytidines* in PAR – CLIP. For HITS – CLIP protocol, *uridine* mutations (to any of A, C, G nucleotides) and deletions were the most frequent mutations, followed by insertions to either side of *uridine*

experiments, CLIP experiments yield a wealth of data about specific RBPs, but the computational resources that are necessary for analyzing deep sequencing data in order to infer the binding sites and the binding specificity of the protein of interest, are not generally available in experimental laboratories.

We have developed the CLIPZ[2] analysis environment that supports the automatic functional annotation of short reads resulting primarily from crosslinking and immunoprecipitation experiments (CLIP). The functional annotation could also be applied to short reads resulting from other types of experiments such as mRNA-Seq, Digital Gene Expression, small RNA cloning, etc. The CLIPZ platform enables visualization and mining of individual data sets as well as analysis involving multiple experimental data sets. Our platform can support collaborative projects involving multiple users and groups of users as well as public and private datasets.

MIRZA, *A biophysical model for inferring microRNA-target site interactions from Argonaute crosslinking and immunoprecipitation data*

Researchers have been frustrated that they had no quantitative model to study the interaction of microRNAs and their targets. At the current stage of knowledge it does not seem surprising to find examples of functional non-canonical microRNA sites involved in degradation and/or translation inhibition of their targets. Nonetheless, many microRNA target prediction methods do not pay much attention to non-canonical sites. This is probably because these sites are particularly difficult to be identified.

In this work, we took advantage of experimentally determined microRNA targets to infer an empirical model (called MIRZA[3]) of microRNA-target interaction from large experimental data sets. Briefly, the likelihood of observing the microRNA sequence bound to mRNA binding site with a specific hybridization structure is calculated.

We inferred its parameters within a *bayesian* probabilistic framework from CLIP data. The inferred parameters largely confirm previous knowledge of microRNA-target interaction and further provide the means to identify functional target sites that are *non-canonical* and would have been difficult to accurately predict by other methods.

We verified the inferred non-canonical binding of various microRNAs using experimental data based on transcriptome-wide measurements of mRNA stability upon over-expression of several microRNAs. We saw a milder effect for non-canonical sites [4] compared to typical canonical sites. On the other hands, because of their abundance, we hypothesized that these sites could contribute to fine tuning the

[2] http://www.clipz.unibas.ch

[3] Source code available at: http://www.mirz.unibas.ch/software.php

[4] predicted by MIRZA

regulation of mRNA stability in specific conditions. Furthermore, in applying MIRZA we discovered that the set of target mRNAs of a given microRNA depends on the microRNAâs expression level and the fraction of non-canonical targets increases with expression level.

6.2 FUTURE DIRECTIONS

The work presented here tackles the challenges of RNA binding proteins of various aspects ranging from bioinformatics for in depth analysis of deep sequencing data, to provide software platform to facilitate the data analysis and finally to use the data in order to infer empirical models for better understanding the regulatory network.

I think the CLIP based methods could still be improved both in terms of experimental procedures and in terms of data analysis. There are still a lot of RNA binding proteins to which one could apply the method in order to identify the binding sites. To answer how the spatio-temporal activity of RBPs post-transcriptionally affects the regulation of gene expression, we still need to think more in terms of both developing novel experimental techniques and also advance data analysis and empirical modeling.

For CLIPZ, the aim is to permanently maintain and constantly develop new tools to provide faster, easier-to-use and more robust services. There is, for example, a lot of potential modifications to the annotation pipeline to minimize the number of mis-annotated reads or developing a novel mapping algorithm which is faster and require less computational resources will improve annotation pipeline efficacy. We plan to include other organisms (i.e. model organisms) in the annotation pipeline in order to make it possible to perform analysis on the high-throughput data obtained from different organisms.

On the aspect of software architecture of the CLIPZ web server, there are still a lot of work to do in order to keep the server as dependable as possible.

Other suggestions could be to provide fault tolerant analysis environment which enables the system to continue its intended operation, possibly at a reduced level, rather than failing completely, when some part of the system fails. We could simplify the CLIPZ software architecture and its dependencies to in order to facilitate end-user customization for local software installations.

MIRZA provides a fundamental model of microRNA-target interaction which represents a unique tool for identification of targets of individual microRNAs and could predict the binding strength of any microRNA to any given mRNA target site. This makes it a valuable tool for the analysis of Argonaute-CLIP data. MIRZA revisits the question of target prediction based on a probabilistic framework, it is straight forward to apply such method to other RNA-target interactions such as snoRNAs.

In the current development state, MIRZA predicts target sites only by using the sequence information of the CLIPed sites. However combining it with other features such as evolutionary conservation improve its predictive power.

Parameter optimization process is implemented such that it make use of parallel computing whenever is needed. However there is still a room for improvement of this process or use other efficient optimization techniques in order to speed up the process. The object oriented architecture of MIRZA algorithm makes it easy to customize the various processing modules (i.e. parameter optimization module) or use it as a component in combination with other algorithms.

Part V

APPENDIX

A SUPPLEMENTARY MATERIAL TO CHAPTER ON PAR-CLIP

A.1 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.

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
```

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 (In-

[1] -NH2 indicates C6 aminolinker (Dharmacon).

vitrogen) 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. miRNAs from immunoprecipitates and the lysate were cloned and Solexa-sequenced [66] 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 purified 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 [4]. 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 unbound 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 [66]. 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.

A.2 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. 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 [14]) 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 [3] and to a database of sequences whose function (rRNA, tRNA, sn/snoRNA, miRNA, mRNA, etc.) is already known. These were obtained from

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

[3] http://genome.cse.ucsc.edu

the sources specified in Berninger et al. The Oligomap algorithm [14] 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 [194] 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 20B,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 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 2E, 3E, and 4E). 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 21. 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 [47]. 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 [181] 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 21). 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 21). 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 [160]. 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.). 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 [47, 53], 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 [180]. To identify binding sites in PUM2 clusters of aligned sequence reads using the inferred weight matrix, we used the MotEvo algorithm [181], 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 20A). 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 [178], 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 ourselves 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[4]. 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 22, blue traces). The binding site was defined to

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

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 [51]. The transcript probe set intensities were background-corrected, adjusted for non-specific binding and quantile normalized with the GCRMA algorithm [195]. 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 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 26). 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 26) 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 26). 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 5C.

The miRNA annotation was preformed as described in Berninger et al., Landgraf et al.

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 [14]. To determine whether the enriched motifs may correspond to miRNAs, all significantly enriched motifs (with a posterior $\geqslant$ 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 [104], 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 procedure using 20 randomly selected miRNAs (Figure 25F) and secondly counting of the number of seed match pair occurrence in the extended clusters for 100 sets of 20 randomly selected miRNAs (Figure 25H). A visualization of seed match pair occurrence is shown in Figure 25G.

Properties of crosslinked and non-crosslinked miRNA seed matches

For the analyses whose results are presented in Figure 26, 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.

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 [61]. 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 [76] 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) [158] 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 [168]. 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 [21], fitting the distance between the seed matches and the STOP codon to the CDS length (Figure 26M,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 26K-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.

A.3 Supplementary tables

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

A.4 Supplementary figures

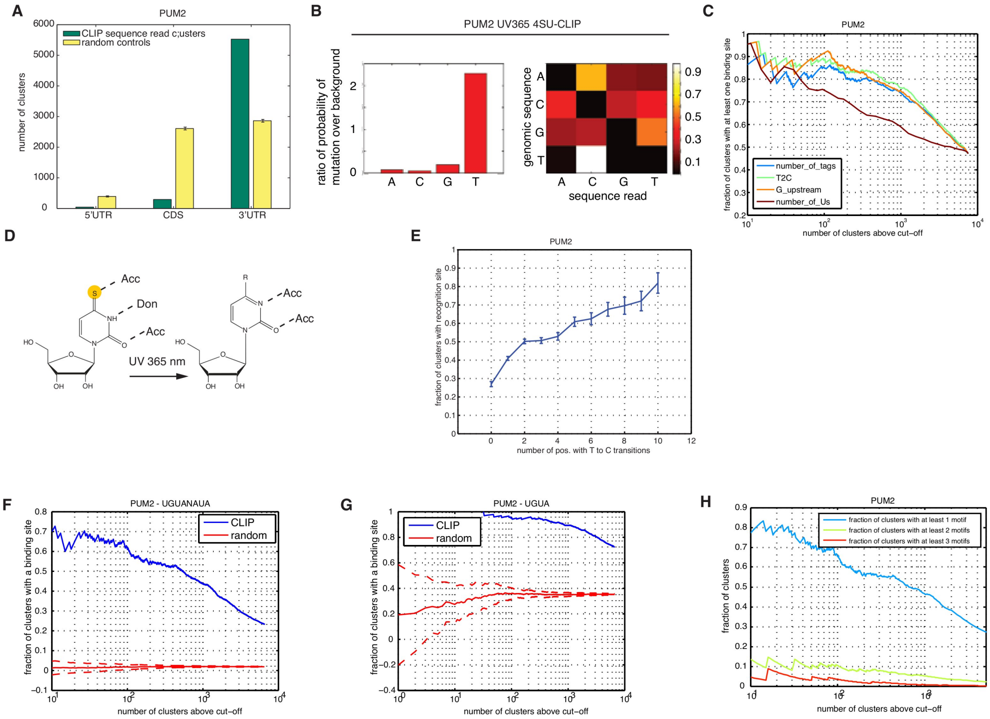


Figure 20: Analysis of PUM2-PAR-CLIP clusters. Related to Figure 2. (A) Analysis of the transcript regional preferences and the mutational pattern of crosslinked sequences of PUM2. The number of exonic sequence read clusters annotated as derived from the 5'UTR, CDS or 3'UTR of a target transcript is 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. (F-H) Enrichment of binding motifs for PUM2 for the consensus motif UGUANAUA

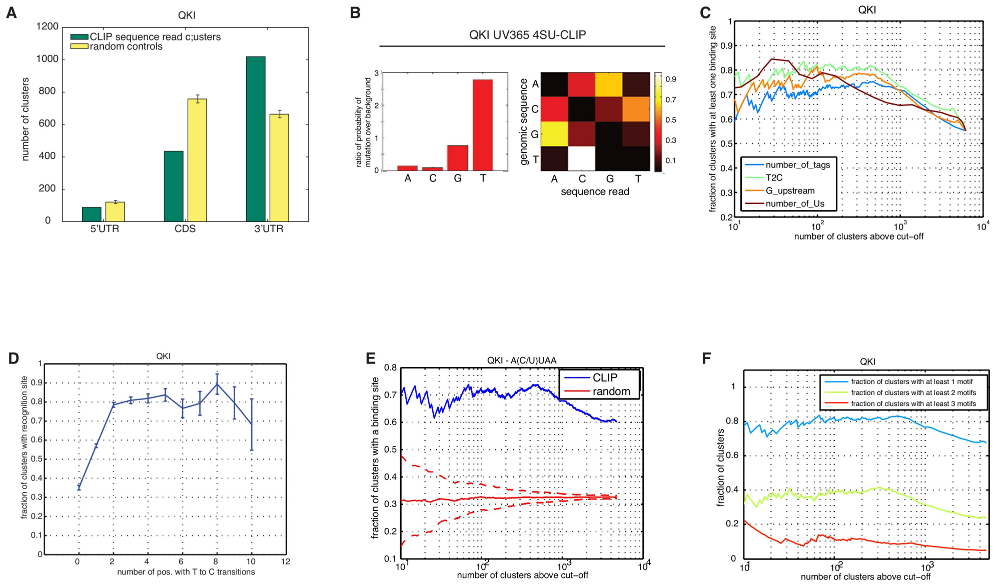


Figure 21: Analysis of QKI-PAR-CLIP clusters. Related to Figure 22. (A) Analysis of the transcript regional preferences and the mutational pattern of crosslinked sequences of QKI. The number of exonic sequence read clusters annotated as derived from the 5'UTR, CDS or 3'UTR of a target transcript is shown (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.

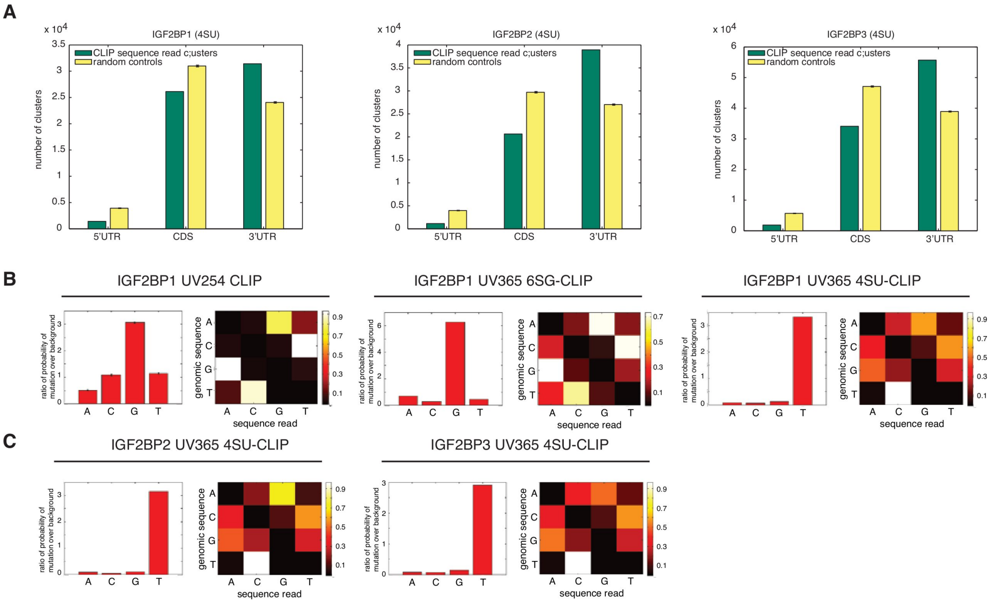


Figure 22: Analysis of IGF2BP1-3-PAR-CLIP clusters. Related to Figure 4. (A) Analysis of the transcript regional preferences and the mutational pattern of crosslinked sequences of IGF2BP1-3. The number of exonic sequence read clusters annotated as derived from the 5'UTR, CDS or 3'UTR of a target transcript 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.

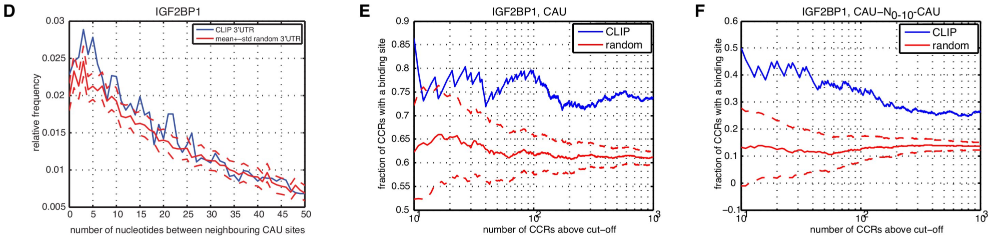
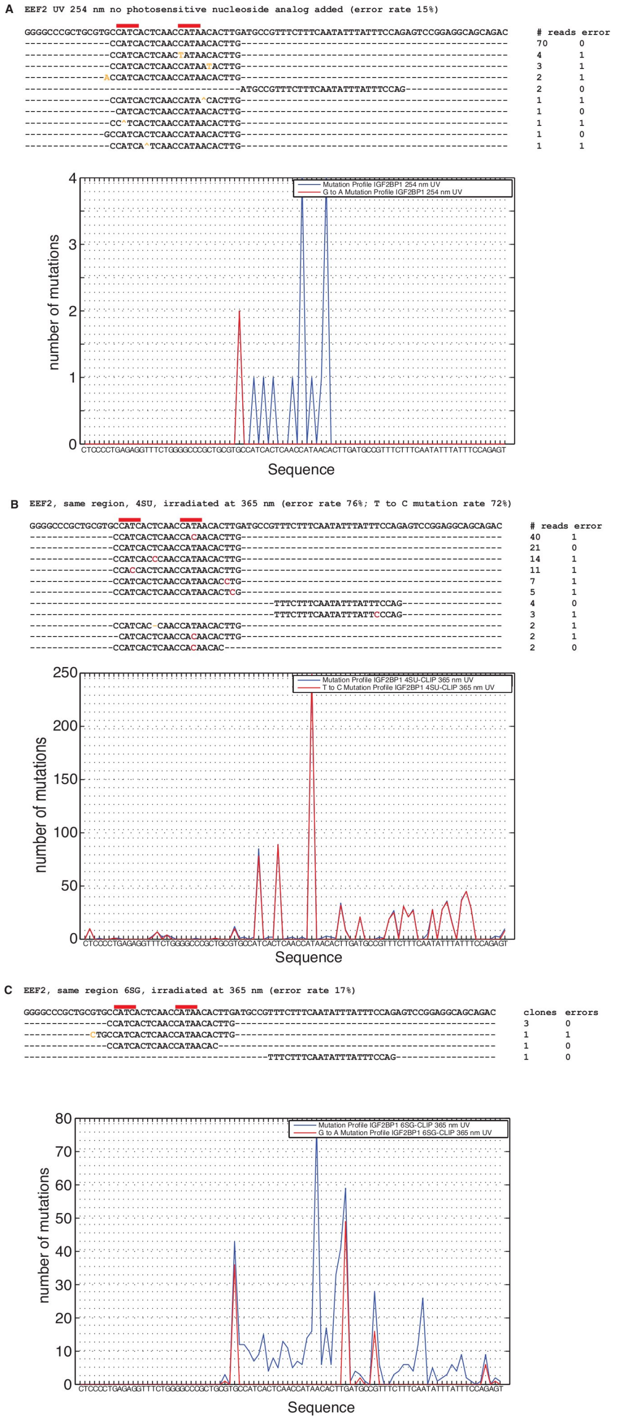


Figure 22: (D) Distance between two neighboring CAU-motifs in crosslinked IGF2BP1 PAR-CLIP clusters (blue line) and in randomized transcripts (red line). CAU-motifs are enriched within 3-5 nt distance of each other in the crosslinked regions compared to randomized sequence sets. Only IGF2BP1 is shown 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).



A EEF2 UV 254 nm no photosensitive nucleoside analog added (error rate 15%)

```
GGGGCCCGCTGCGTGCCATCACTCAACCATAACACTTGATGCCGTTTCTTTCAATATTTATTTCCAGAGTCCGGAGGCAGCAGAC     # reads error
---------------CCATCACTCAACCATAACACTTG-----------------------------------------------     70      0
---------------CCATCACTCAACTATAACACTTG-----------------------------------------------     4       1
---------------CCATCACTCAACCATAATACTTG-----------------------------------------------     3       1
--------------ACCATCACTCAACCATAACACTTG-----------------------------------------------     2       1
--------------------------------------ATGCCGTTTCTTTCAATATTTATTTCCAG------------------     2       0
---------------CCATCACTCAACCATA^CACTTG-----------------------------------------------     1       1
----------------CATCACTCAACCATAACACTTG-----------------------------------------------     1       0
---------------CC^TCACTCAACCATAACACTTG-----------------------------------------------     1       1
--------------GCCATCACTCAACCATAACACTTG-----------------------------------------------     1       0
---------------CCATCA^TCAACCATAACACTTG-----------------------------------------------     1       1
```

4
3
2
1
0
number of mutations
Mutation Profile IGF2BP1 254 nm UV
G to A Mutation Profile IGF2BP1 254 nm UV
CTCCCCTGAGAGGTTTCTGGGGCCCGCTGCGTGCCATCACTCAACCATAACACTTGATGCCGTTTCTTTCAATATTTATTTCCAGAGT
Sequence

B EEF2, same region, 4SU, irradiated at 365 nm (error rate 76%; T to C mutation rate 72%)

```
GGGGCCCGCTGCGTGCCATCACTCAACCATAACACTTGATGCCGTTTCTTTCAATATTTATTTCCAGAGTCCGGAGGCAGCAGAC     # reads error
---------------CCATCACTCAACCACAACACTTG-----------------------------------------------     40      1
---------------CCATCACTCAACCATAACACTTG-----------------------------------------------     21      0
---------------CCATCACCCAACCATAACACTTG-----------------------------------------------     14      1
---------------CCACCACTCAACCATAACACTTG-----------------------------------------------     11      1
---------------CCATCACTCAACCATAACACCTG-----------------------------------------------     7       1
---------------CCATCACTCAACCATAACACTCG-----------------------------------------------     5       1
--------------------------------------------TTTCTTTCAATATTTATTTCCAG------------------     4       0
--------------------------------------------TTTCTTTCAATATTTATTCCCAG------------------     3       1
---------------CCATCAC-CAACCATAACACTTG-----------------------------------------------     2       1
----------------CATCACTCAACCACAACACTTG-----------------------------------------------     2       1
---------------CCATCACTCAACCACAACAC--------------------------------------------------     2       0
```

250
200
150
100
50
0
number of mutations
Mutation Profile IGF2BP1 4SU-CLIP 365 nm UV
T to C Mutation Profile IGF2BP1 4SU-CLIP 365 nm UV
CTCCCCTGAGAGGTTTCTGGGGCCCGCTGCGTGCCATCACTCAACCATAACACTTGATGCCGTTTCTTTCAATATTTATTTCCAGAGT
Sequence

C EEF2, same region 6SG, irradiated at 365 nm (error rate 17%)

```
GGGGCCCGCTGCGTGCCATCACTCAACCATAACACTTGATGCCGTTTCTTTCAATATTTATTTCCAGAGTCCGGAGGCAGCAGAC     clones  errors
---------------CCATCACTCAACCATAACACTTG-----------------------------------------------     3       0
------------CTGCCATCACTCAACCATAACACTTG-----------------------------------------------     1       1
---------------CCATCACTCAACCATAACAC--------------------------------------------------     1       0
--------------------------------------------TTTCTTTCAATATTTATTTCCAG------------------     1       0
```

80
70
60
50
40
30
20
10
0
number of mutations
Mutation Profile IGF2BP1 6SG-CLIP 365 nm UV
G to A Mutation Profile IGF2BP1 6SG-CLIP 365 nm UV
CTCCCCTGAGAGGTTTCTGGGGCCCGCTGCGTGCCATCACTCAACCATAACACTTGATGCCGTTTCTTTCAATATTTATTTCCAGAGT
Sequence

Figure 23: 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 4. 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).

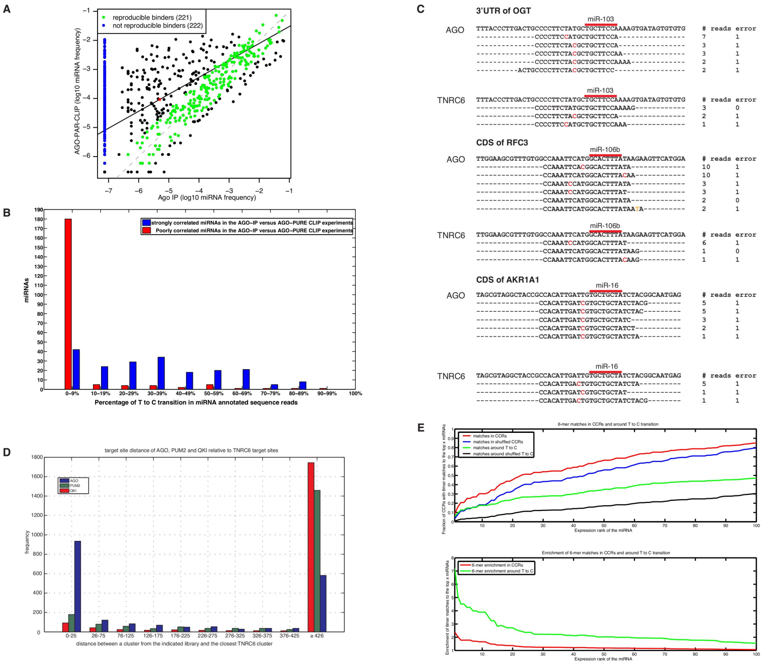


Figure 24: AGO-protein family PAR-CLIP. Related to Figure 5. (A) Principal component analysis of the relative abundance of miRNAs derived from the combination of the AGO-PAR-CLIP libraries on one hand, and the non-crosslinked AGO-IPs on the other hand. The first principal component is projected onto 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).

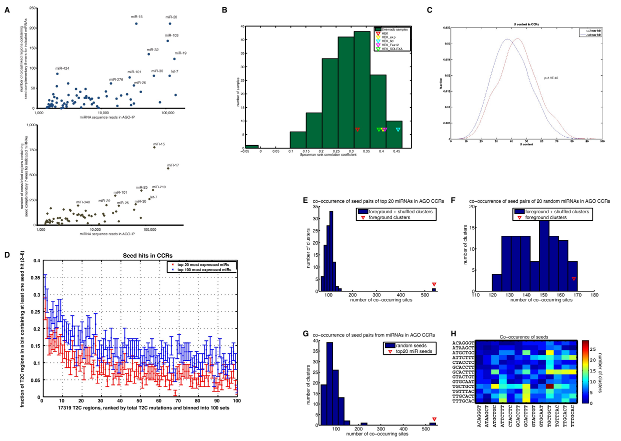


Figure 25: Seed complementary sequences from abundant HEK293 miRNAs are enriched in AGO-PAR-CLIP CCRs. Related to Figure 6. CCRs from the AGO-PAR-CLIP are enriched for target sites for the most abundant miRNAs in HEK293 cells. (A) Correlation between occurrence of 8-mer (upper panel) and 7-mer (lower 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. (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.

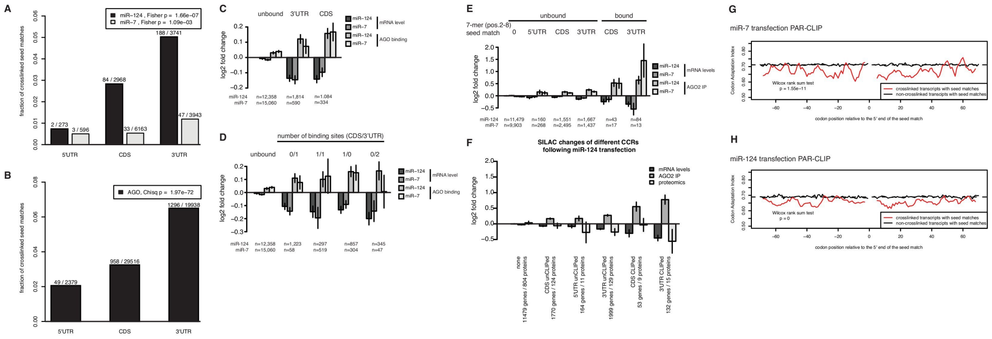


Figure 26: Properties of CCRs containing miRNA seed complementary sites. Related to Figure 7. (A) Seed complementary sequences in the 3′UTR are more efficiently crosslinked than seed complementary regions in the CDS. Fraction of crosslinked seed matches (1-7 or 2-8) for the miR-124 (dark bars) and miR-7 (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 3 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 [73]. 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.) 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.

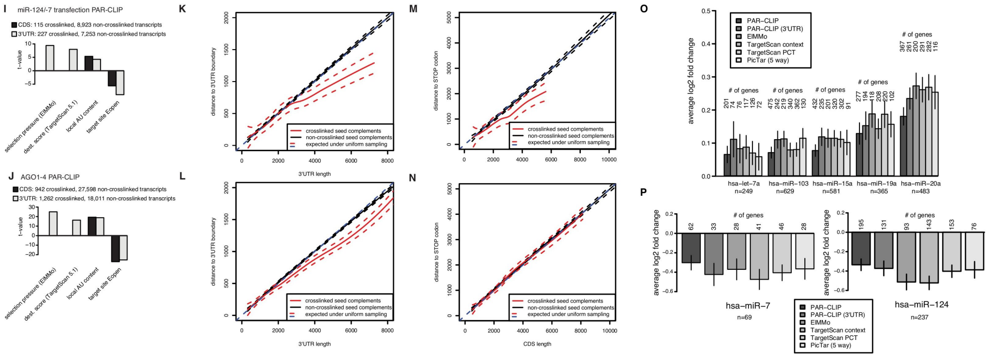
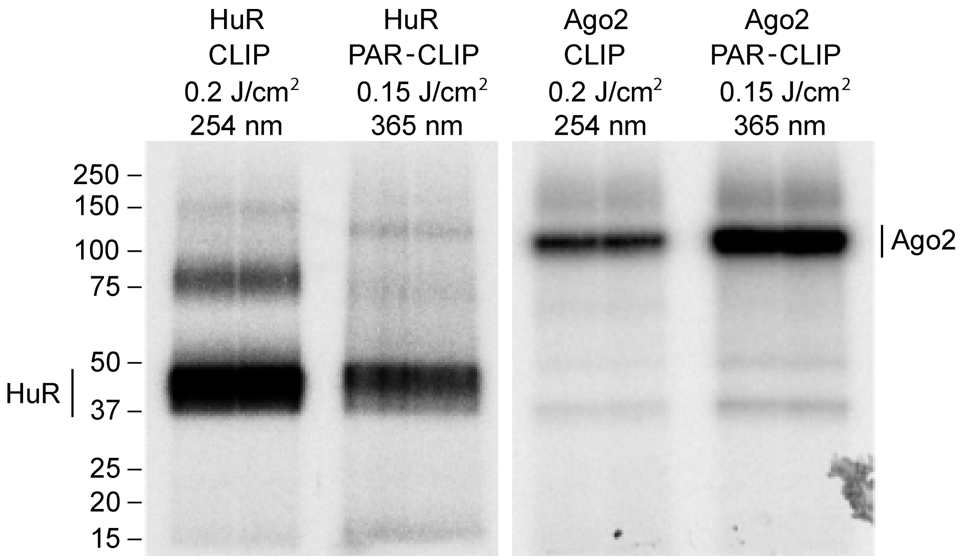


Figure 26: (I) The sequence context defines a functional miRNA binding site in the UTR as well as in the CDS. Four different criteria (selection pressure, destabilization score, local A/U content, target site openness) were compared for crosslinked transcripts containing 7-mer seed matches for a miR-124 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.

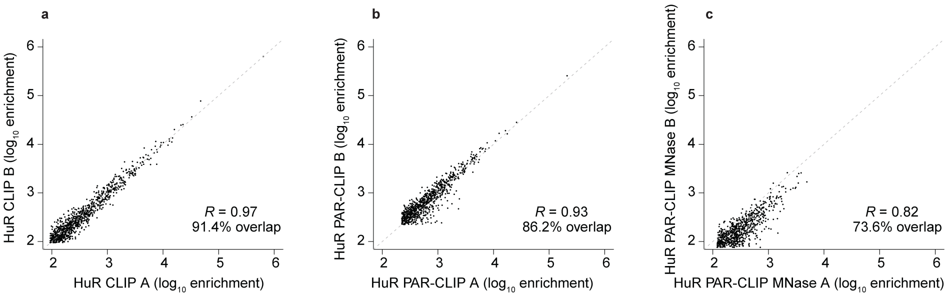
B

SUPPLEMENTARY MATERIAL TO THE CHAPTER ON QUANTITATIVE ANALYSIS OF CLIP METHODS

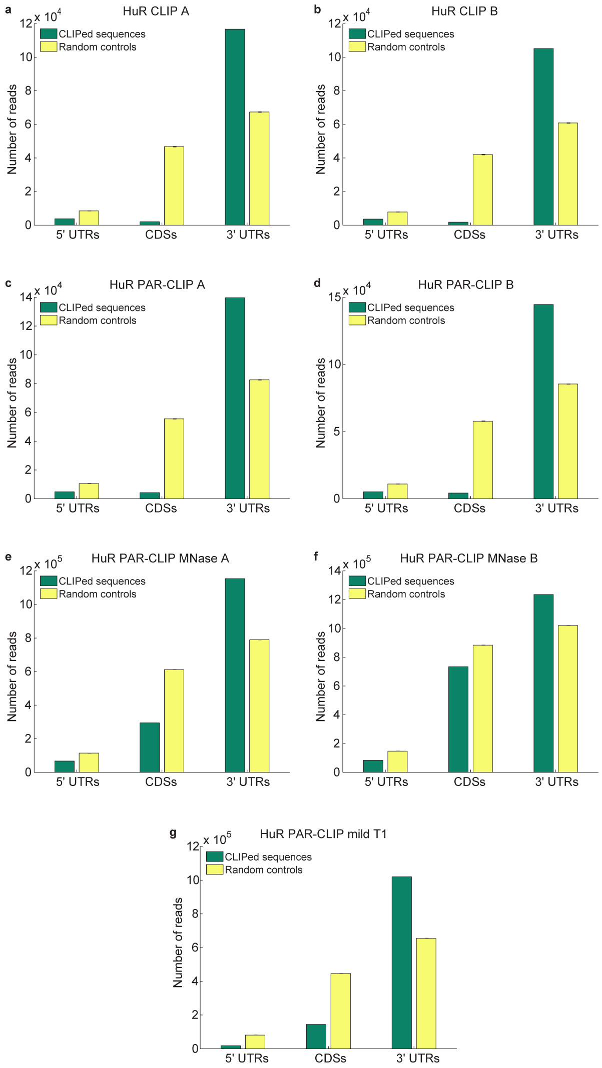
B.1 SUPPLEMENTARY FIGURES



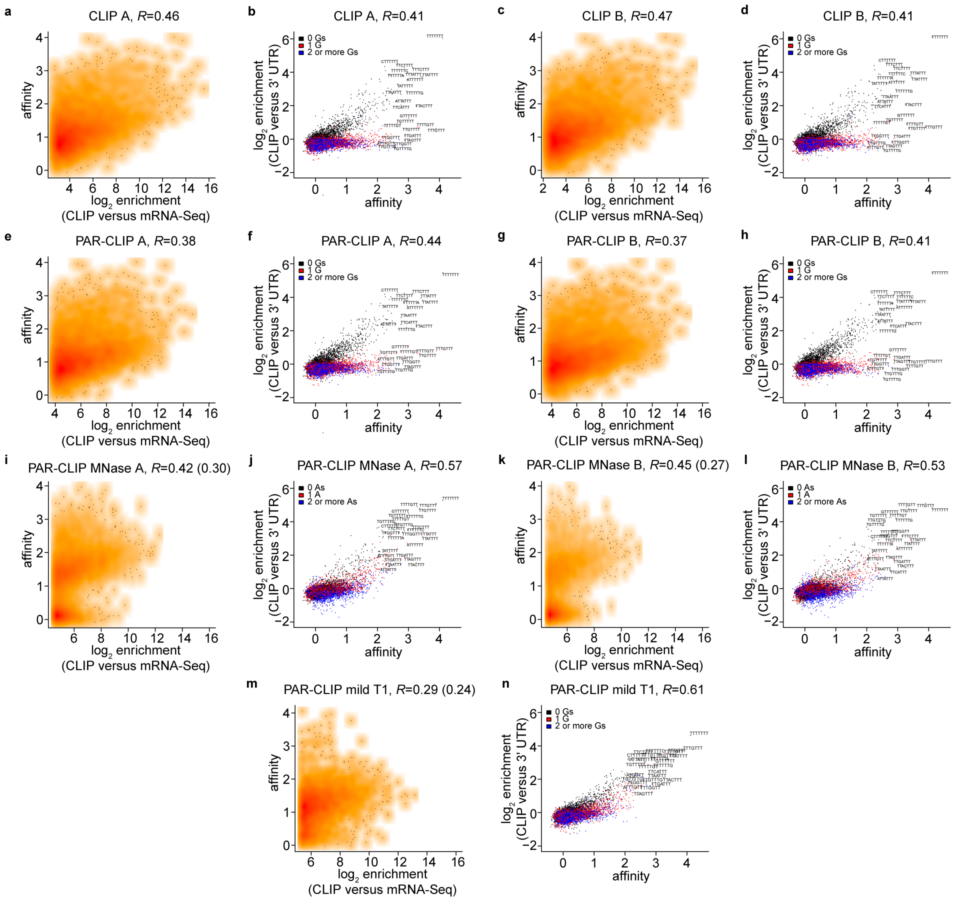
Suppl. Fig. 27: Crosslinking efficiency of CLIP and PAR-CLIP. Autoradiograph of the protein gel after IP.



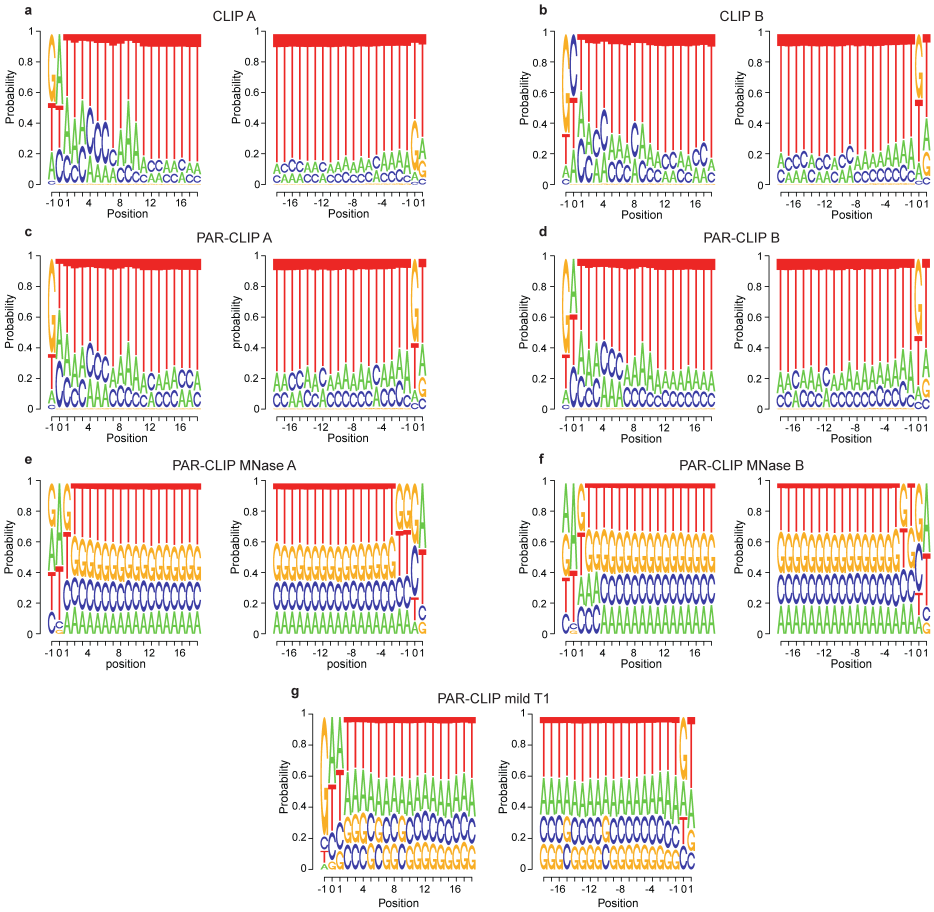
Suppl. Fig. 28: Correlation between the enrichment in reads in individual HuR sites among CLIP (a), PAR-CLIP (b) and PAR-CLIP MNase (c) replicate experiments. Each point on the plot represents an individual binding site. The correlation coefficient and the fraction of the 1000 most enriched sites that overlap between replicate experiments are indicated.



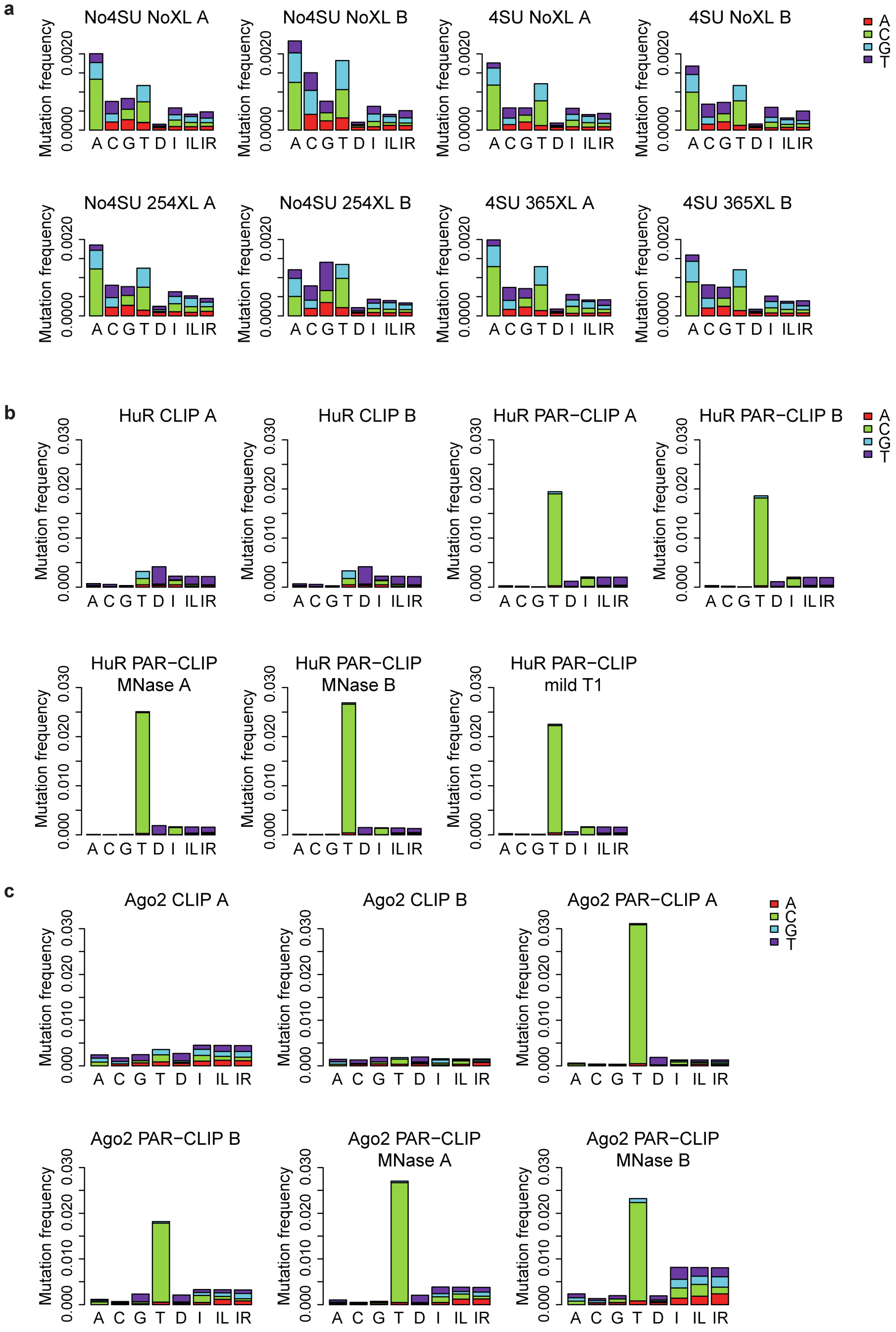
Suppl. Fig. 29: Contour plots of the distribution of enrichment, relative to mRNA abundance, of sequence reads in HuR binding sites determined with CLIP (a,c), PAR-CLIP (e,g), PAR-CLIP MNase (i,k) and PAR-CLIP mild T1 (m) and the predicted affinity of the sites for HuR determined based on RNAcompete data [147]. Correlation coefficients outside the brackets for panels (i,k,m) indicate were calculated only based on the points in the higher cloud. Correlations between the estimated affinity of a 7-mer motif and its enrichment in CLIP (b,d), PAR-CLIP (f,h), PAR-CLIP MNase (j,l) and PAR-CLIP mild T1 (n) binding sites relative to 3′ UTRs.



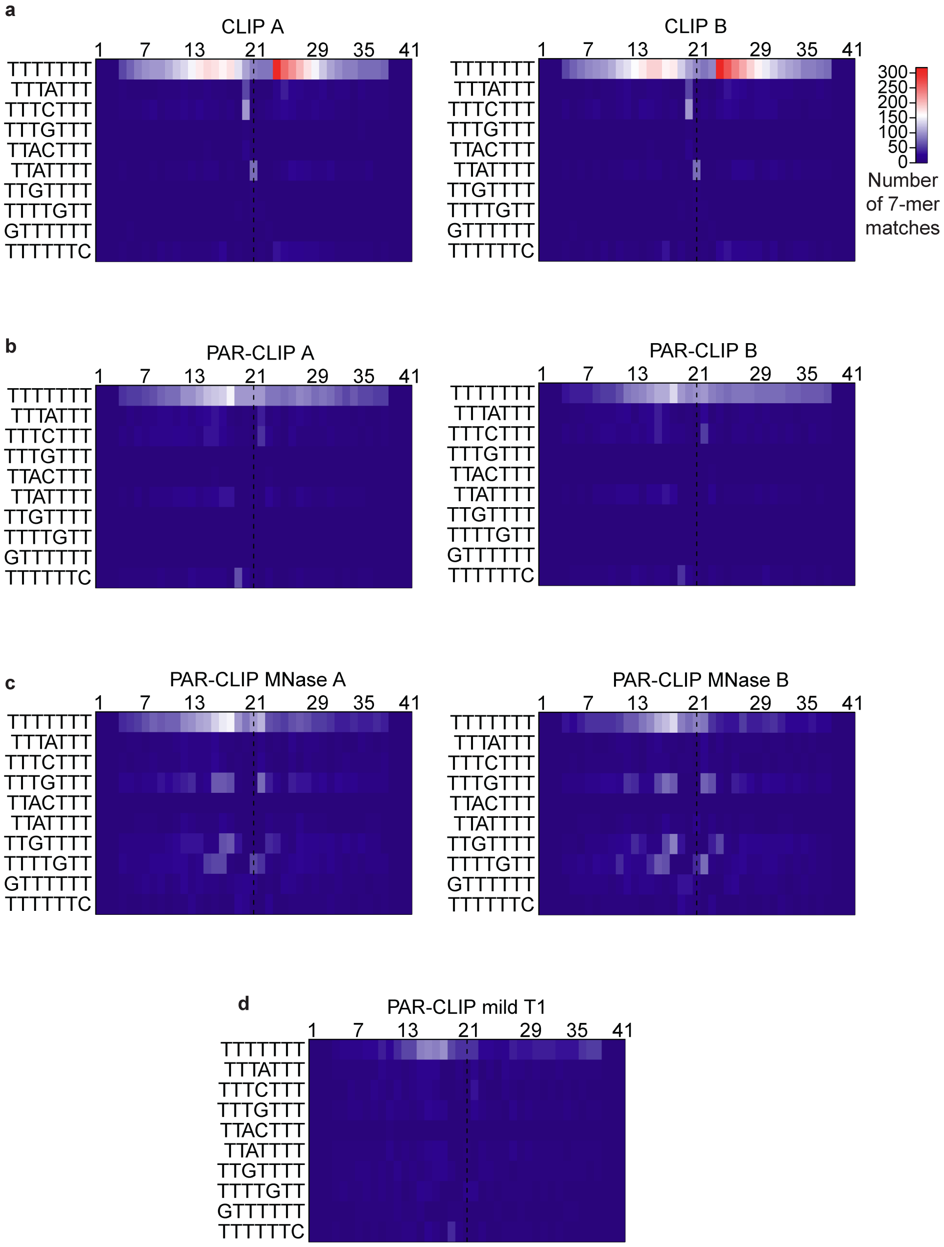
Suppl. Fig. 30: Mutation bias in mRNA-seq (a), CLIP, PAR-CLIP and PAR-CLIP MNase reads obtained for HuR (b) and Ago2 (c) proteins. We determined the frequency of various types of mutations (substitution, deletion, insertion) in mRNA-annotated reads that mapped with at most one error to the genome. The first four columns in each plot correspond to substitutions (the frequencies of substitutions towards each of the three possible nucleotides are indicated by the different colors), the fifth column to deletions relative to the genome sequence (deletions of the four nucleotides are also indicated separately) and the sixth column to insertions relative to the genome sequence (inserted nucleotides shown separately). The seventh and eighth columns show the identity of nucleotides that are located 5′ (IL) and 3′ (IR) of an inserted nucleotide. The sample names are indicated on the panels.



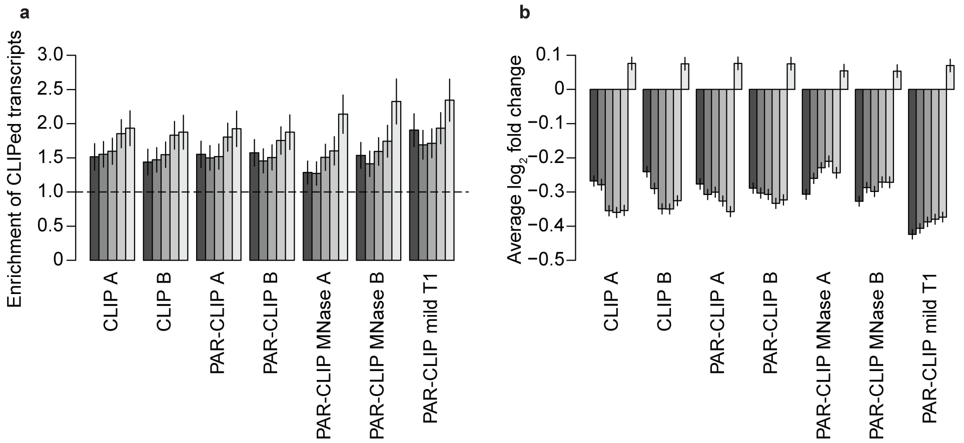
Suppl. Fig. 31: Observed (green) distribution of HuR reads between 5′ UTR, CDS and 3′ UTR regions of transcripts, and the expected distribution (yellow) based on the relative length of these regions in the transcripts from which the reads originated.



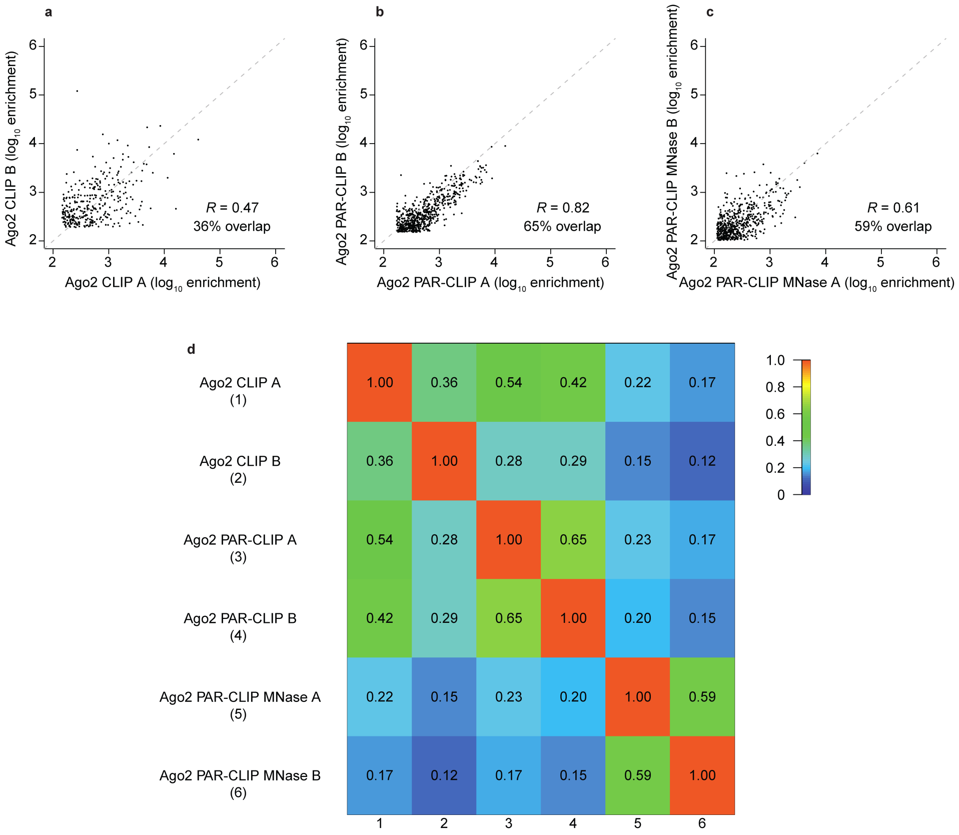
Suppl. Fig. 32: Position-wise nucleotide frequencies in reads obtained in HuR CLIP (a, b), PAR-CLIP (c, d), PAR-CLIP MNase (e, f) and PAR-CLIP mild T1 (g) samples. Reads were anchored either at the 5′ (left-hand side plot of each set) or the 3′ (right-hand side plot of each set) end. The location of the 5′ and 3′ ends corresponds to position 0. Positions upstream of the anchor end are labeled with negative numbers and positions downstream of the anchor end with positive numbers.



Suppl. Fig. 33: Location of the ten 7-mers with highest affinity for HuR within CLIP (a), PAR-CLIP (b), PAR-CLIP MNase (c) and PAR-CLIP mild T1 crosslink-centered regions. The central positions of the 7-mers are used to compute the heatmap. The position of the predominant mutation (T deletion or mutation to G/A/C in CLIP and T-to-C in PAR-CLIP) is indicated by a dashed line.



Suppl. Fig. 34: (a) Enrichment of CLIPed transcripts (i.e. transcripts that contain at least one of the top 5000 CLIPed sites) among all significantly downregulated transcripts upon HuR siRNA transfection. Transcripts were divided into non-overlapping bins from least to most expressed in the GFP siRNA samples. The enrichment in CLIPed HuR targets among all the downregulated transcripts was then computed separately for each individual bin. Errorbars denote standard error of the mean (b) Mean expression change upon HuR knockdown of transcripts carrying the top 1000, 1001-2000, ... , 4001-5000 sites for HuR, as determined by various CLIP methods. Binding sites were sorted based on their enrichment, divided into non-overlapping bins of 1000 sites, and the change in expression of the host transcripts upon HuR knockdown was computed. The last bin for each sample shows the average fold-change of all the expressed transcripts that did not contain any of the top 5000 sites. Errorbars denote standard error of the mean

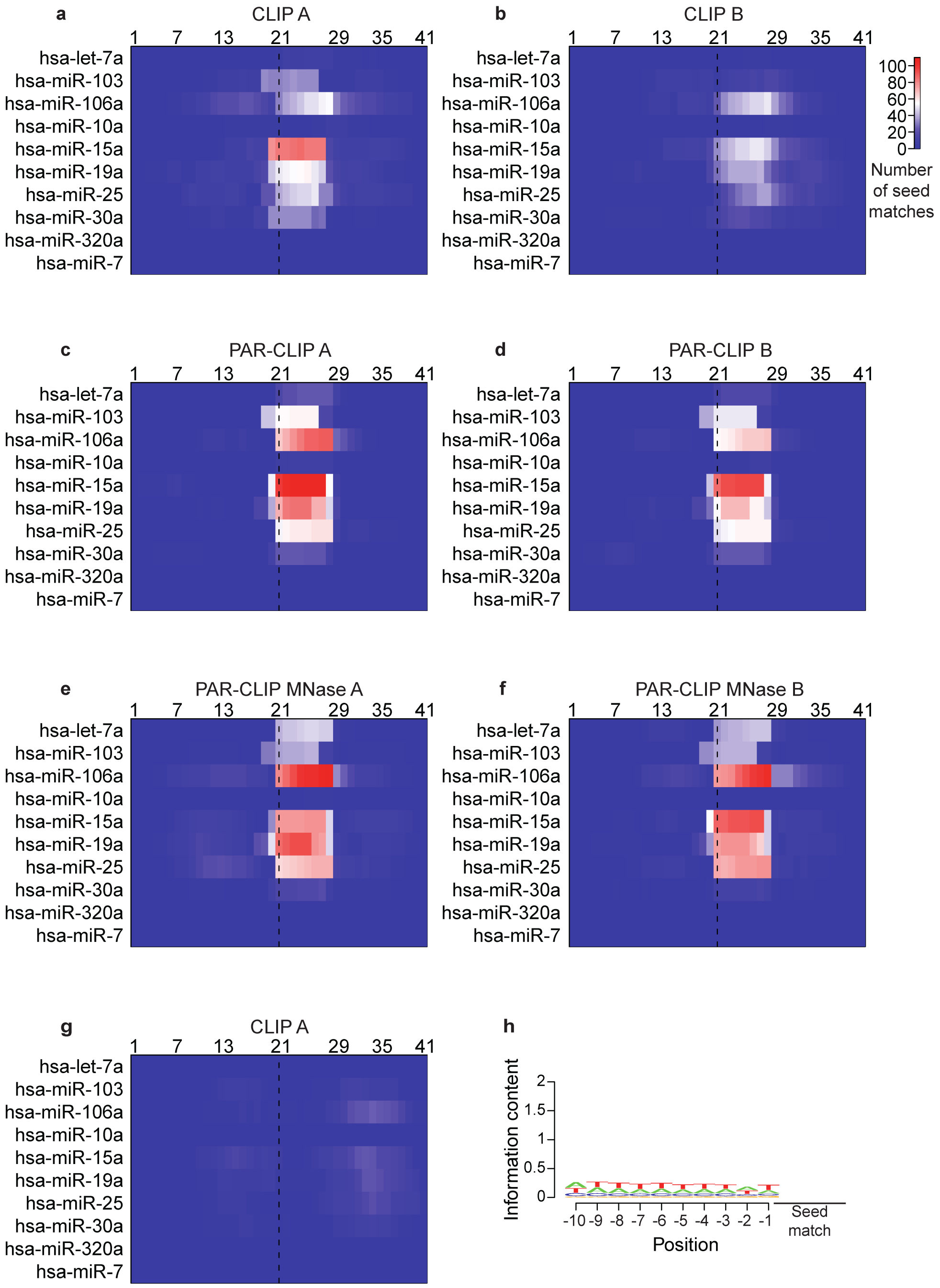


	1	2	3	4	5	6
Ago2 CLIP A (1)	1.00	0.36	0.54	0.42	0.22	0.17
Ago2 CLIP B (2)	0.36	1.00	0.28	0.29	0.15	0.12
Ago2 PAR-CLIP A (3)	0.54	0.28	1.00	0.65	0.23	0.17
Ago2 PAR-CLIP B (4)	0.42	0.29	0.65	1.00	0.20	0.15
Ago2 PAR-CLIP MNase A (5)	0.22	0.15	0.23	0.20	1.00	0.59
Ago2 PAR-CLIP MNase B (6)	0.17	0.12	0.17	0.15	0.59	1.00

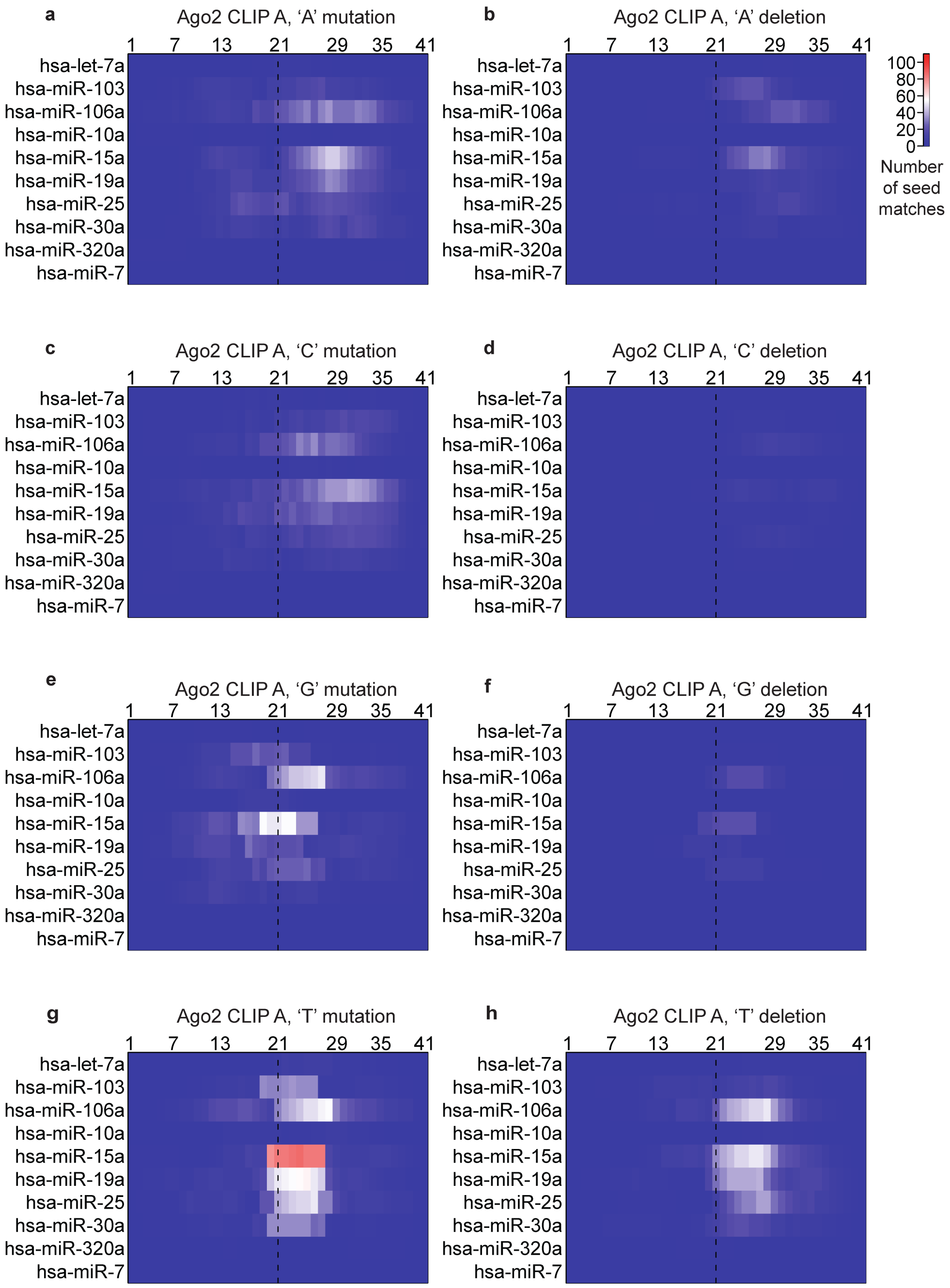
Suppl. Fig. 35: Western blot showing HuR downregulation upon siRNA transfection. After detection with the HuR antibody, the blot was reprobed for hnRNP C. The asterisks (*) mark antibody cross-reactivity.

b.2 supplementary tables

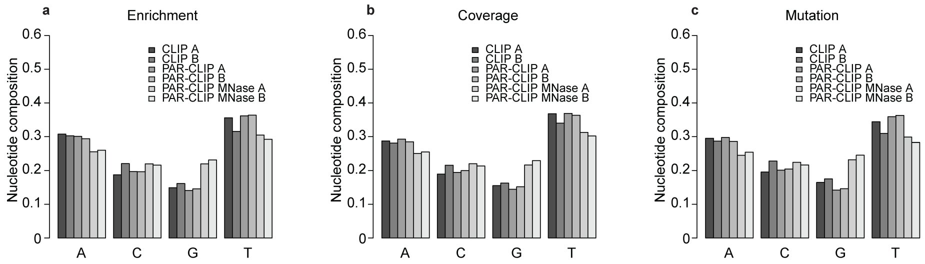
Supplementary table S1 is available in the online supplementary material of [95].



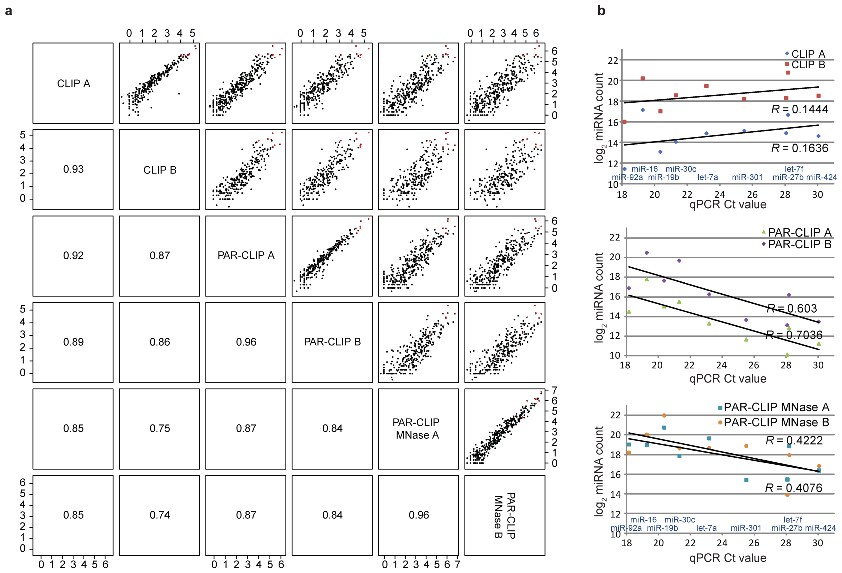
Suppl. Fig. 36: Reproducibility of Ago2 binding site identification in CLIP (a), PAR-CLIP (b) and PAR-CLIP MNase (c) experiments. From each sample we selected the top 1000 binding sites according to the enrichment in reads in the site relative to the mRNA abundance. We computed the proportion of sites that are identified in any given pair of samples (d), not only for the replicates. For the replicates we also computed the correlation between enrichment values in the two experiments.



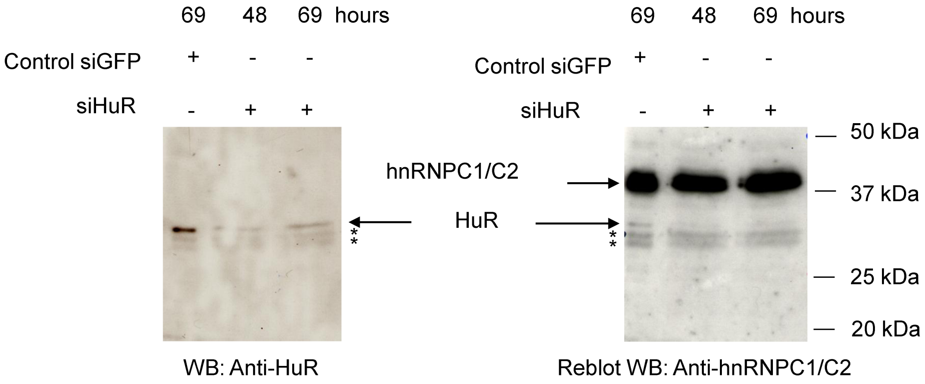
Suppl. Fig. 37: (a) Reproducibility of miRNA profiles constructed based on various types of CLIP experiments. The x- and y-axes indicate the $\log_{10}$ counts of a given miRNA in a pair of samples. (b) Correlation between the miRNA expression level in total RNA (expressed as multiplication cycles in RT-PCR (Ct values) and the expression level in the CLIP samples (expressed as $\log_2$(read counts)). The miRNAs that were measured in this experiment are indicated by red dots on panel (a).



Suppl. Fig. 38: Location of matches to the seed of the ten most abundant miRNA families relative to the position of the most abundant mutation of a particular type (which is located in the center of the the 41-nucleotide long regions) in the 1000 most enriched Ago2 sites. As crosslink-diagnostic mutation we took T mutation or T deletion in the CLIP-samples (panels a and b), and T-to-C mutation in PAR-CLIP samples (panels c, d, e and f). Panel g shows a similar heatmap constructed based on CLIP A sample and considering only insertions as crosslink-diagnostic mutations. Panel h shows the relative frequency of the four nucleotides immediately upstream of the seed match among CLIP A sites.



Suppl. Fig. 39: Location of matches to the seed of the ten most abundant miRNA families relative to the position of the most abundant mutation of a particular type (which is located in the center of the the 41-nucleotide long regions) in the 1000 most enriched Ago2 CLIP A sites. The mutation type is specified in the title of each plot.



Suppl. Fig. 40: Base composition of Ago2 sites obtained with different protocols. The pattern is similar whether the sites are extracted based on the enrichment in reads (a), coverage by reads (b) or density of crosslink-diagnostic mutations (defined as in Suppl. Fig37) (c).

C SUPPLEMENTARY MATERIAL TO CHAPTER MIRZA: A BIOPHYSICAL MODEL FOR INFERRING MICRORNA-TARGET SITE INTERACTIONS FROM ARGONAUTE CROSSLINKING AND IMMUNOPRECIPITATION DATA

C.1 TESTS OF THE INFERENCE PROCEDURE WITH SYNTHETIC DATA

To test the ability of our algorithm to correctly infer MIRZA's energy parameters, we designed two synthetic data sets. First, we constructed the reverse complement of positions 1 through 8 (the "seed" region) of 10 selected miRNAs. We called these sequences miRNA seed matches. For each miRNA we then generated 300 target sequences by embedding the miRNA seed match at a random position in a sequence of length 40, in which all other nucleotides were chosen with uniform probability from the 4 possible nucleotides. We thus generated a pool of 3000 synthetic mRNA target fragments corresponding to the prototypical, miRNA seed-matching sites [116]. We inferred the parameters of the MIRZA model from this input data set through simulated annealing. To conservatively estimate the accuracy of the parameter inference, we repeated the simulated annealing 25 times and recorded the variation in the inferred parameters across the runs.

As shown in Suppl. Fig. 41A, the inferred biophysical parameters correctly reflect that only positions $1-8$ contribute to the binding of miRNA and target in this synthetic data set, the inferred energies of A-U and C-G base pairings being essentially identical, as expected. Binding at other positions and internal loops are strongly disfavored. Furthermore, the fitted vector of target fractions $\vec{\pi_\mu}$ contained values of approximately 0.1 for all miRNAs (results not shown), correctly reflecting the fact that we engineered an equal number of target mRNA fragments for each miRNA. The target qualities of the $3'000$ correct combinations of mRNA fragment m and miRNA μ were consistently higher (Suppl. Fig. 41B, in green) than those of all $27'000$ incorrect combinations of m and μ (Suppl. Fig.41B in red) demonstrating that on this simple synthetic data set the algorithm had perfect performance.

In a second test, we asked our algorithm to infer a more complex interaction model from a synthetic pool of mRNA fragments constructed as follows:

1. The target mRNA fragments, 300 for each of 10 selected miRNAs, were longer (50 nucleotides). Nucleotide frequencies were not equal, but rather 0.3 for A and U and 0.2 for G and C nucleotides.

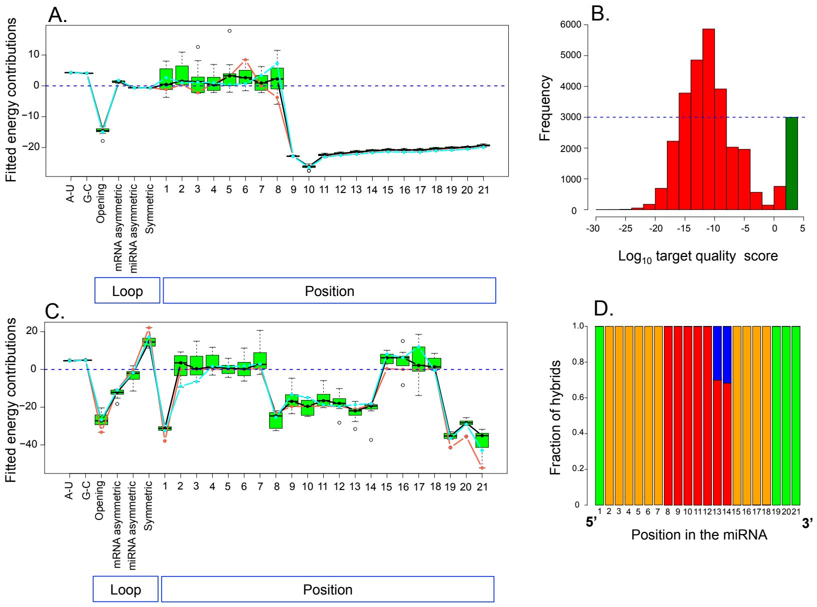


Figure 41: Binding models inferred from synthetic data. **A:** Summary of model parameter values inferred from the synthetic data set with seed-type interactions only. *Green* boxes indicate inter-quartile ranges and whiskers indicate the 5 and 95 percentiles across the 25 simulated annealing runs. Parameters of the models that yielded the highest and second-highest probability of the data are shown in *red* and *blue*, respectively, and median values of fitted parameters across the runs are show in *black*. **B:** Histogram of target quality scores for all miRNA-mRNA fragment pairs. The model that gave the highest probability of the data was used to compute target quality scores $R(m|\mu)$ of each mRNA fragment m in the synthetic seed-type data set with each of the 10 miRNAs μ used in the test. Scores for the *correct* and *incorrect* miRNA-mRNA associations are shown in *green* and *red*. **C:** Summary of the model parameter values inferred from the synthetic data set with sites that have 3′-compensatory-type interactions. Box-plots summarize parameter values fitted across 10 independent optimization runs as in panel A. **D:** Statistical summary of the most likely hybrid structures for the synthetic data-set with 3′-compensatory-type sites. Each column corresponds to a miRNA position and the colors show the fractions of optimal hybrids that had different structural features at the corresponding position. *Green:* dangling end nucleotides. *Orange:* hybridized nucleotides. *Red:* nucleotides involved in symmetrical loops. *Blue:* miRNA nucleotides that are bulged out.

2. Embedded at a random location in these mRNA fragments was a miRNA-matching region composed of a contiguous match to positions $2-7$ of the miRNA, followed by a loop of $5-7$ nucleotides in the mRNA, and another contiguous match to positions $15-18$ of the miRNA. The length of the loop was chosen uniformly randomly in the range $[5, 7]$.

This type of miRNA-binding sites is similar to the 3′-compensatory sites previously described in the literature [18]. For this more complex data set as well, our algorithm correctly identified positive contributions of base-pairings involving nucleotides $2-7$ and $15-18$ of the miRNA, and negative contributions from all other positions (Suppl. Fig. 41C). Besides disfavouring the opening of loops in general, the algorithm inferred a higher penalty for bulges involving mRNA nucleotides compared to bulges involving miRNA nucleotides, correctly reflecting that the region in between the two engineered miRNA-mRNA helices was generally shorter in the mRNA than in the miRNA ($5-7$ vs. 7 nucleotides).

By identifying, for each mRNA fragment m, the hybrid structure with the highest target frequency, we determined the frequencies with which different hybridization states (see Methods) were used at each position of the miRNA (Suppl. Fig. 41D). The results indicate that the model perfectly reconstructs the hybrid structures, i.e. positions $2-7$ and $15-18$ were always correctly inferred to be paired, while the central region of the miRNA was inferred to be looped out. Because in the mRNA we introduced loops of $5-7$ nucleotides while in the miRNA there were always 7 looped out nucleotides, the algorithm also inferred correctly that positions $8-12$ of the miRNA should be part of symmetrical loops, while positions $13-14$ should either be part of symmetrical loops (when the loop introduced in the mRNA was 7 nucleotides long) or should be asymmetrically looped out (when the loop introduced in the mRNA was 5 or 6 nucleotides in length). The algorithm also inferred correctly the cognate miRNA for each of the 3′000 target sites (data not shown). In summary, MIRZA also obtained perfect performance for this more realistic synthetic data-set.

C.2 FITS TO AGO2 CLIP DATA RESULT IN HIGHLY REPRODUCIBLY PREDICTED TARGET QUALITIES

To fit energy parameters for real miRNA target sites, we performed 100 simulated annealing runs on the Ago2 CLIP data, starting from different initial conditions. As shown in Fig. 18B of the main article, these fits yielded very similar, yet not identical sets of parameters. To further test the robustness of the inferred miRNA targets with different parameters settings we calculated, for the 5 parameter sets with the highest likelihoods, the target qualities $R(m|\mu)$ for all pairs of mRNA fragment m and miRNA μ. We then computed correlation coefficients between the target qualities predicted by different parameter sets and found that they were all very close to 1 (the minimum correlation coefficient, that was obtained for the models with the highest and lowest likelihoods among the 5, was still 0.96, Suppl. Fig. 42A). This indicates that the parameters obtained in different runs yield highly similar miRNA target predictions.

C.3 MIRNA-TARGET SITE INTERACTIONS INFERRED BASED ON RNA-RNA HYBRIDIZATION

We used RNAduplex [124] to predict, for each of the 2′988 target site fragments of our Ago2 CLIP data set, the miRNA that would form the most stable interaction with the fragment, and the structure of the miRNA-mRNA hybrid for this miRNA. Based on these predictions, we inferred the frequency with which different positions in the miRNA and mRNA are hybridized. We found that the hybrids predicted by RNAduplex are very different from those inferred by MIRZA, particularly that miRNA nucleotides are involved in base-pairing with very similar frequencies, irrespective of their position (Suppl. Fig. 42B).

C.4 INFERRED ABUNDANCE OF MIRNAS IN RISC CORRELATES WITH THEIR EXPRESSION

We correlated the fitted miRNA prior, reflecting the relative abundance of a miRNA in RISC, with the proportion of reads assigned to that miRNAs from all miRNA-annotated reads obtained in CLIP experiments. We downloaded the miRNA expression profiles of the 6 Ago2-CLIP samples from Kishore et al. [95] from the CLIPZ web server (`http://www/clipz.unibas.ch`) and computed the average proportion of each miRNA across these 6 samples. The correlation between miRNA expression and the inferred miRNA prior is shown in Suppl. Fig. 42D.

C.5 NON-CANONICAL TARGET SITES ARE EVOLUTIONARILY CONSERVED

As an additional test of functionality of predicted non-canonical sites, we investigated whether these sites may be under evolutionary selection for interacting with the miRNAs. To obtain a set of non-canonical sites we constructed the distribution of target frequencies $R(m|\mu)\pi_\mu$ for all canonical sites identified in the Ago2 CLIP data, and obtained its mean and standard deviation. We then selected all non-canonical sites with a target frequency up to half a standard-deviation below the mean target frequency of canonical sites. There were 77 such high scoring non-canonical sites in the Ago2 CLIP data set.

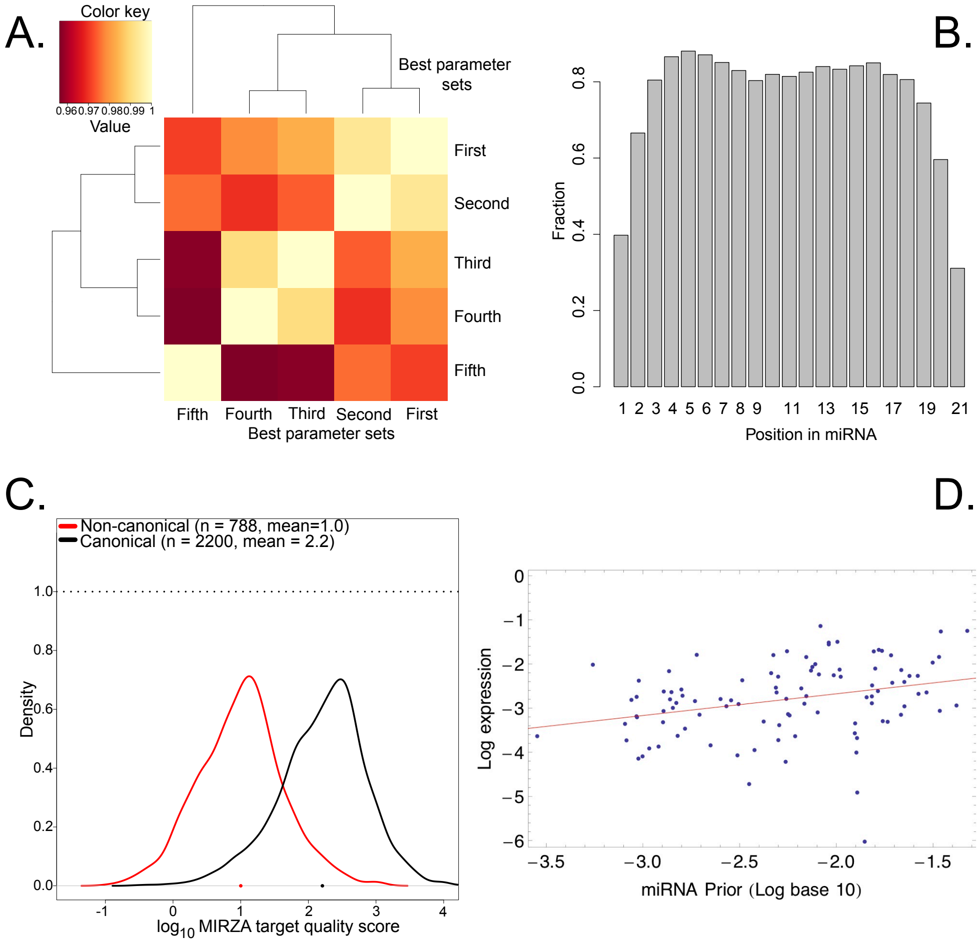
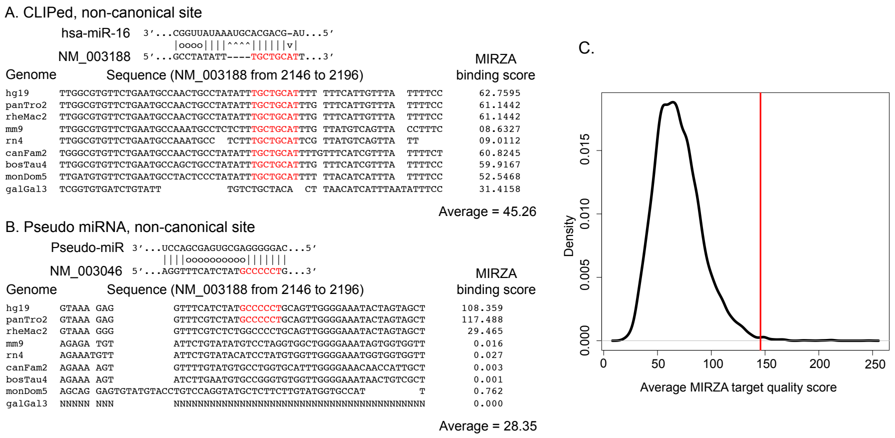


Figure 42: Results of the parameter inference based on Ago2-CLIP data. **A:** Pairwise correlations between the target qualities $R(m|\mu)$ obtained for each mRNA fragment - miRNA pair (m, μ) across the 5 models from the Argonaute 2 CLIP data in independent simulated annealing runs. **B:** Structure of the hybrids predicted by RNAduplex [124]. Each column corresponds to a position in the miRNA and shows the fraction of the best hybrids (as predicted by RNAduplex) in which the corresponding position of the miRNA was hybridized. **C:** Probability densities of target quality scores $R(m|\mu)$ for Ago2 CLIP sites that were predicted to form canonical (*black*) and non-canonical (*red*) best hybrids with the miRNAs that yielded the highest target quality score. **D:** Scatter plot of the miRNA expression levels (proportion of reads associated with a given miRNA in Ago2-CLIP data among all miRNA-annotated reads in the CLIP data) against the inferred prior for the corresponding miRNAs. All 6 Ago2-CLIP data set from Kishore et al. [95] were used to compute average miRNA expression level. Expression profiles were extracted from the CLIPZ web server (http://www.clipz.unibas.ch). The Spearman correlation coefficient was 0.36 (P-value = 2.3×10^{-4}).

For each predicted non-canonical target site we extracted its genomic location in the reference genome [1] and extracted orthologous regions from other vertebrate genomes [2]. With the MIRZA model, we then determined the target quality $R(m|\mu)$ of each of the orthologous sites and calculated the average target quality across all orthologous sites.



A. CLIPed, non-canonical site

```
hsa-miR-16  3'...CGGUUAUAAAUGCACGACG-AU...5'
                  |oooo||||^^^^||||||v|
NM_003188   5'...GCCTATATT----TGCTGCATT...3'
```

```
Genome    Sequence (NM_003188 from 2146 to 2196)                    MIRZA
                                                                    binding score
hg19      TTGGCGTGTTCTGAATGCCAACTGCCTATATTTGCTGCATTTT TTTCATTGTTTA  TTTTCC       62.7595
panTro2   TTGGCGTGTTCTGAATGCCAACTGCCTATATTTGCTGCATTTG TTTCATTGTTTA  TTTTCC       61.1442
rheMac2   TTGGCGTGTTCTGAATGCCAACTGCCTATATTTGCTGCATTTG TTTCATTGTTTA  TTTTCC       61.1442
mm9       TTGGCGTGTTCTGAATGCCAAATGCCTCTCTTTGCTGCATTTG TTATGTCAGTTA  CCTTTC       08.6327
rn4       TTGGCGTGTTCTGAATGCCAAATGCC  TCTTTGCTGCATTCG TTATGTCAGTTA  TT           09.0112
canFam2   TGGGCGTGTTCTGAATGCCAACTGCCTATATTTGCTGCATTTTGTTTCATCGTTTA  TTTTCT       60.8245
bosTau4   TTGGCGTGTTCTGAATGCCAGCTGCCTATATTTGCTGCATTTG TTTCATCGTTTA  TTTTCC       59.9167
monDom5   TTGATGTGTTCTGAATGCCTACTCCCTATATTTGCTGCATTTT TTACATCATTTA  TTTTCC       52.5468
galGal3   TCGGTGTGATCTGTATT           TGTCTGCTACA  CT TAACATCATTTAATATTTCC       31.4158

                                                                    Average = 45.26
```

B. Pseudo miRNA, non-canonical site

```
Pseudo-miR  3'...UCCAGCGAGUGCGAGGGGGAC...5'
                  ||||oooooooooo|||||||
NM_003046   5'...AGGTTTCATCTATGCCCCCTG...3'
```

```
Genome    Sequence (NM_003188 from 2146 to 2196)                    MIRZA
                                                                    binding score
hg19      GTAAA GAG          GTTTCATCTATGCCCCCTGCAGTTGGGGAAATACTAGTAGCT     108.359
panTro2   GTAAA GAG          GTTTCGTCTATGCCCCCTGCAGTTGGGGAAATACTAGTAGCT     117.488
rheMac2   GTAAA GGG          GTTTCGTCTCTGGCCCCTGCAGTTGGGGAAATACTAGTAGCT      29.465
mm9       AGAGA TGT          ATTCTGTATATGTCCTAGGTGGCTGGGGAAATAGTGGTGGTT       0.016
rn4       AGAAATGTT          ATTCTGTATACATCCTATGTGGTTGGGGAAATGGTGGTGGTT       0.027
canFam2   AGAAA AGT          GTTTTGTATGTGCCTGGTGCATTTGGGGAAACAACCATTGCT       0.003
bosTau4   AGAAA AGT          ATCTTGAATGTGCCGGGTGTGGTTGGGGAAATAACTGTCGCT       0.001
monDom5   AGCAG GAGTGTATGTACCTGTCCAGGTATGCTCTTCTTGTATGGTGCCAT         T       0.762
galGal3   NNNNN NNN          NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN       0.000

                                                                    Average = 28.35
```

Figure 43: Evaluation of evolutionary conservation of non-canonical sites. **A:** A high-scoring non-canonical site inferred from the Argonaute 2 CLIP data is shown at the top of the panel. The mRNA subsequence that hybridizes with the miRNA seed region is highlighted in red. The alignment block of the genomic region in which this site resides in nine vertebrate species is shown underneath the hybrid and the MIRZA target quality score for the interaction of miR-16 with each of the orthologous sites is shown on the right side of the sequence. **B:** An example of a high-scoring non-canonical site complementary to a pseudo-miRNA is shown at the top of panel. The mRNA subsequence that hybridizes with the seed region of the pseudo-miRNA is marked in red. Shown are also the alignment of orthologous regions and the MIRZA target quality scores for the interaction with these regions. **C':** Distribution of average target quality scores computed over orthologs of non-canonical binding sites predicted by MIRZA for pseudo-miRNAs. *Black:* Distribution of average MIRZA target quality computed over 10′000 re-sampled sets of orthologs of non-canonical binding sites for pseudo-miRNAs. Each re-sampled set had the same number of sites as the set of non-canonical miR-7 CLIPed sites. Averages were taken first over orthologous sequences and then over all sites in the a randomized set. *Red:* Average MIRZA target quality score of the non-canonical miR-7 CLIPed sites and their orthologs.

To determine whether the conservation of the orthologous regions suggests purifying selection for interacting with the miRNA, we compared these target qualities with the target qualities that we obtained

[1] Human (hg19 assembly version from the University of California at Santa Cruz)

[2] Rat (rn4), Mouse (mm9), Chimpanzee (panTro2), Opossum (monDom5), Dog (canFam2), Cow (bosTau4), Rhesus (rheMac2), Chicken (galGal3)

for randomized data sets constructed as follows. We generated a set of *pseudo-miRNAs*, consisting of random 21-mer sequences chosen to preserve the base composition of real miRNAs at each position of the miRNA. We then scanned all human 3′ UTR sequences with a sliding a window of 51 nucleotides, and selected high-scoring non-canonical sites for *pseudo-miRNAs*, with the same threshold on target quality as used for the sites of real miRNAs. We then sampled the same number of non-canonical sites of the pseudo-miRNAs as the number of non-canonical CLIPed sites. We determined orthologs of these sites in exactly the same way as was done for the original CLIPed sites and computed the average target quality of these orthologous sites with MIRZA. The procedure is sketched in Suppl. Fig. 43. We randomly sampled 10′000 times a set of pseudo-targets of the same size as the set of true non-canonical targets and calculated the average target quality of the orthologs of these sites. Supplementary Fig. 43C shows the distribution of average ortholog target qualities for the sets of pseudo-targets (black curve). Only 48 of the 10′000 sets of high scoring non-canonical pseudo-miRNA sites had an average score that was at least as high as the average score of the non-canonical CLIPed sites (shown as a vertical red line), giving us an estimated *p-value* of 0.0048. In summary, the non-canonical target sites predicted by MIRZA show significant evidence of being under purifying selection for retaining their target quality.

C.6 COMPARISON OF TARGET PREDICTION ACCURACY

As detailed in the Methods, we used microarray measurements of mRNA expression changes upon miRNA transfection from 5 data sets [123, 61, 156, 113, 50], covering a total of 38 transfection experiment on 26 different miRNAs, to assess the performance of different target prediction methods. Besides target predictions by MIRZA we obtained lists of predicted targets for 10 other target prediction methods including methods that use conservation (TargetScan Pct, ElMMo, PicTar), methods that use the sequence context of the site (TargetScan context+, MIRANDA), a motif frequency-based method (RNA22) and methods that model the energy of interaction between the miRNA and the target (PITA, RNAhybrid, RNAduplex). In addition, we downloaded lists of predicted targets from the Starbase database [196] which intersects Ago2 CLIP sites with miRNA target predictions by TargetScan, PicTar, Miranda, PITA, and RNA22.

For each method and each transfection data set, we mapped the predicted targets to Entrez genes and sorted the predicted targets by the score assigned by the method (see Methods). We then calculated the *median log fold-change* of the top n predicted targets as a function of n for each method. For Starbase we calculated the median log fold-change for two lists provided on the web site: the ′default′ list, and the most comprehensive list that is obtained with the most lenient cut-offs. Figure 19C of the main text shows the median log fold-change as a function of the number of top predictions for MIRZA, the 9 other methods (colored lines), and the Starbase predictions (gray dots), averaged over all transfection experiments. As described in the online Methods, by comparing the fraction of predicted targets that are downregulated with the fraction expected by chance, we can also estimate the total number of functional targets that is predicted by each method for each transfection experiment. Figure 19D in the main text shows these totals averaged over all transfection experiments.

One of the advantages of MIRZA's biophysical model is that is calculates an affinity between miRNA and target (i.e. target quality) which takes into account the contribution of the seed region without having to explicitly filter for or select sites that obey a particular definition of a 'seed match'. This allows MIRZA to identify both canonical and non-canonical sites using as single biophysical scoring function. We thus also compared the performance of the different methods in identifying functional non-canonical target sites by calculating the exact same performance measures, but now restricting ourselves to all predicted target genes that do not have any canonical target sites (see Online Methods). The performance on non-canonical targets, averaged over all transfection experiments, is shown in Figures 19E and 19F in the main text.

C.6.1 *Performance on individual data sets*

Here we show the performance of MIRZA and the other target prediction methods, separately for each of the 5 data sets. Suppl. Fig. 44A shows the average performance of the methods on the 16 transfection experiments from [123]. Similarly, Suppl. Fig. 44B shows the average performance of the methods on the 9 transfection experiments from [61], Suppl. Fig. 44C the performance on the 4 transfection experiments of [156], Suppl. Fig. 44D the performance on the 7 transfection experiments of [113], and Suppl. Fig. 44E the performance on the 2 transfection experiments of [50].

The average results over all transfection experiments (Fig. 19C-F of the main text) showed that MIRZA's targets show the largest down-regulation and that MIRZA, TargetScan Pct, Targetscan context+, ElMMo, and MIRANDA predict the largest number of functional targets. For the 5 individual data sets we generally see the same behavior; with the exception of the data set of Grimson et al. [61], the extent of down-regulation of MIRZA targets is at least as large as for any other method.

With respect to the total number of predicted targets, although there is some variation across the 5 data sets, the results the individual data sets are also largely consistent with the global average. Namely, when

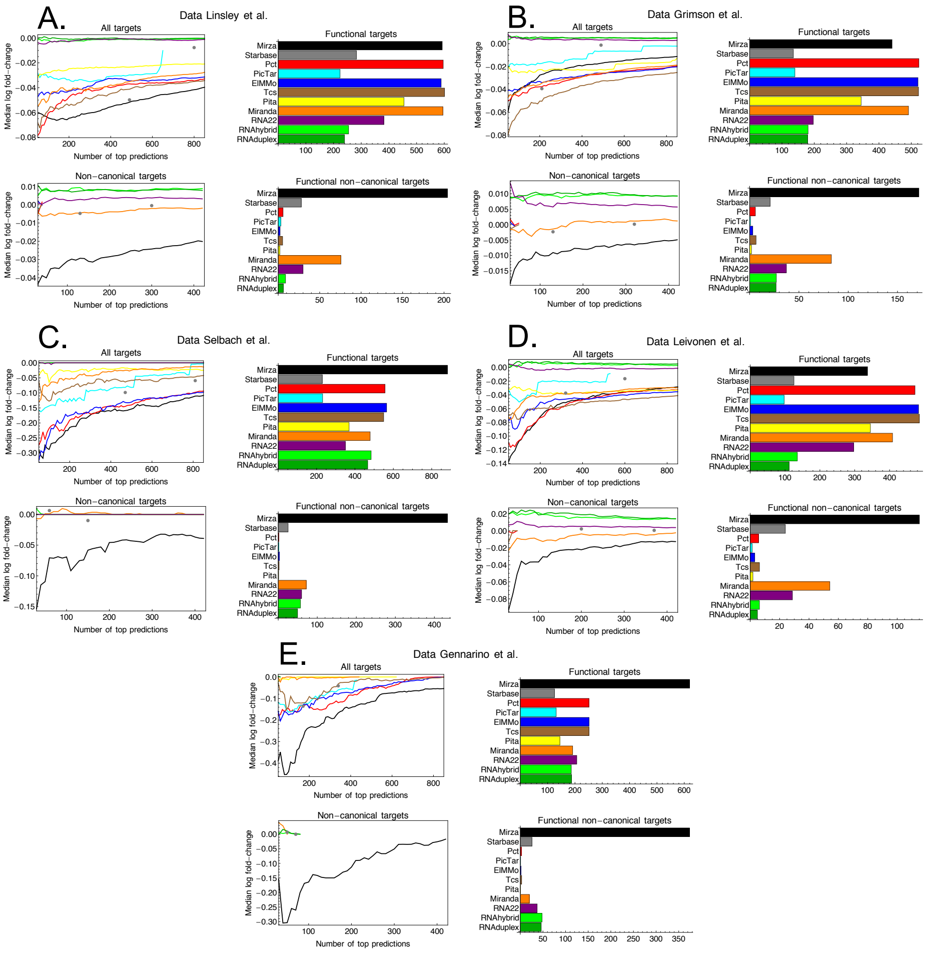


Figure 44: Performance comparison on various individual transfection data sets: **A:** Linsley et al. [123], **B:** Grimson et al. [61], **C:** Selbach et al. [156], **D:** Leivonen et al. [113] and **E:** Gennarino et al. [50]. There are 4 panels for each data set. The upper left panel shows the observed median log-fold changes of target genes, as predicted by several miRNA target prediction methods, averaged over the transfection experiments. For each method the predicted targets were sorted by their target prediction score and the curves show the median log fold-change of the top n targets on the vertical axis as a function of n on the horizontal axis. Each color corresponds to a target prediction method. *Black*: MIRZA, *Red*: TargetScan P_{ct}, *Cyan*: PicTar, *Blue*: ElMMo, *Brown*: TargetScan context+, *Yellow*: PITA, *Orange*: MIRANDA, *Violet*: RNA22, *Light green*: RNAhybrid, *Green*: RNAduplex. Median log fold changes for the lists of targets provided by Starbase are shown as gray dots. The upper right panel shows the estimated number of functional targets predicted by each method as described in the Online Methods, averaged over all transfections in the data set. The methods are indicated next to the bars. The two bottom panels show similar quantities, but now restricted to all predicted target genes that do not have any canonical target site for the corresponding miRNA. Note that the same colors are used to denote the different methods in all panels.

considering all targets MIRZA obtains at least as good a performance as the best performing of the tested methods. Among previously published methods TargetScan variants tend to have the highest performance, followed by ElMMo, PicTar, and MIRANDA. PITA typically performs worse, and the worst performance is observed for RNA22 and methods that simply predict the energy of interaction of small RNAs with target sites, with the standard energy parameters for RNA secondary structure prediction (RNAhybrid and RNAduplex). These observations on the relative accuracy of different methods are consistent with previous evaluations [2, 156]. It is also noteworthy that, by intersecting CLIPed sites with sites predicted by the various target prediction methods, Starbase provides lists of targets that are typically functional, i.e. with good median down-regulation for most data sets. However, because the approach of Starbase is to intersect CLIPed sites with miRNA target predictions, the list of predicted targets in Starbase are generally smaller, leading to overall significantly lower numbers of functional targets.

The most dramatic difference in performance between MIRZA and the other methods is observed when predicting non-canonical targets. We find that, consistently across all 5 data sets, MIRZA's predicted non-canonical targets show much stronger down-regulation and much larger numbers of functional targets than any of the other methods. Furthermore, as is the case with canonical targets, there is a clear correlation between the MIRZA score of non-canonical targets and the degree to which they undergo down-regulation upon miRNA transfection. As mentioned before, many miRNA target prediction methods explicitly require a match to the miRNA, and thus these methods do not even appear in the panels with median fold-changes on non-canonical targets. Of the methods that do predict a substantial number of non-canonical targets, RNA22, RNAhybrid and RNAduplex invariably perform poorly, typically predicting only a small number of functional non-canonical targets. MIRANDA appears able to identify some functional non-canonical targets but it is strongly outperformed by MIRZA on all data sets. In summary, our comparison shows that MIRZA is the only method that can reliable identify a substantial number of functional non-canonical targets.

c.6.2 *Performance comparison on individual miRNAs*

We also compared the performance on the 8 individual miRNAs from the data-set of Linsley et al. [123], both for all predicted targets (Suppl. Fig. 45) and non-canonical targets (Suppl. Fig. 46). Here we averaged the results over the two transfection experiments (done in two different cell lines) for each miRNA.

As shown in Suppl. Fig. 45, for most of the tested miRNAs, MIRZA's predicted targets show the largest median down-regulation of all methods. The only counter example is miR-200a, where MIRZA is outperformed by other methods, presumably because this miRNA has low expression in HEK293 cells and its targets are not CLIPed in this cell type.

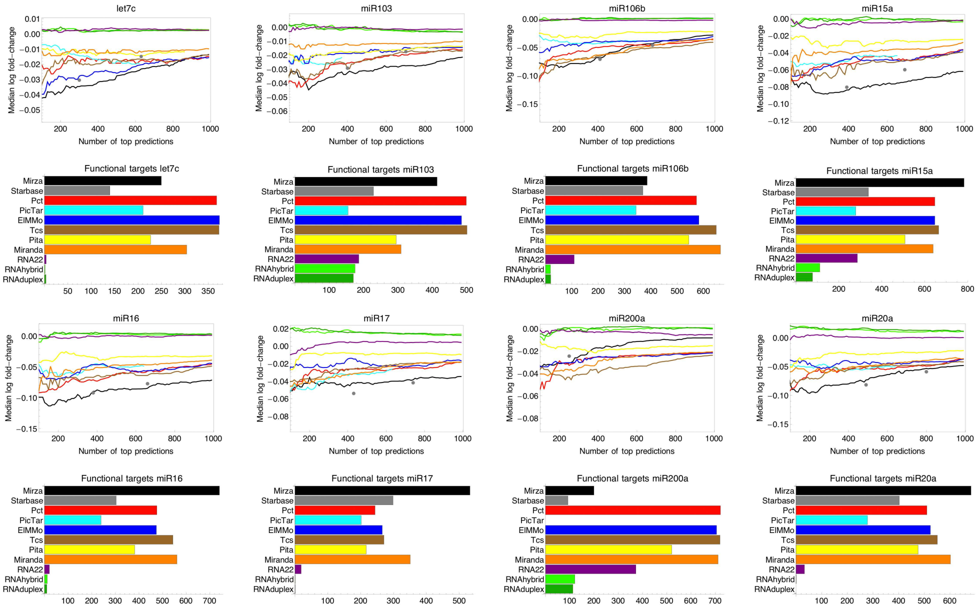


Figure 45: **Top and third row of panels:** Observed median log-fold changes of target transcripts, as predicted by several miRNA target prediction methods, under transfection of the corresponding miRNA (experimental data from [123]). Each panel corresponds to one miRNA transfection, with the transfected miRNA indicated at the top of the panel. For each method the predicted targets were sorted by their target prediction score and the curves show the median log fold-change of the top n targets on the vertical axis as a function of n on the horizontal axis. Each color corresponds to a target prediction method. *Black*: MIRZA, *Red*: TargetScan P_{ct}, *Cyan*: PicTar, *Blue*: ElMMo, *Brown*: TargetScan context+, *Yellow*: PITA, *Orange*: MIRANDA, *Violet*: RNA22, *Light green*: RNAhybrid, *Green*: RNAduplex. Median log fold changes for the lists of targets provided by Starbase are shown as gray dots. **Second and bottom row of panels:** For each miRNA we also estimated the total number of functional targets predicted by each method as described in the Online Methods. The panels show the total number of functional targets predicted for each miRNA (one panel per miRNA) and each method (colored bars). The methods are indicated next to the bars. Note that the same colors are used to denote the different methods in all panels.

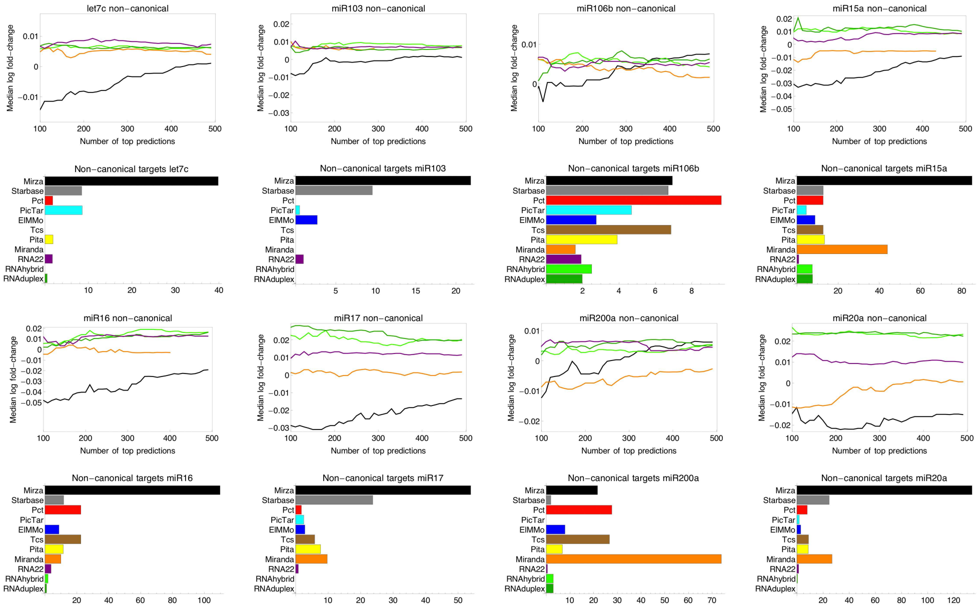


Figure 46: As in Suppl. Fig. 45 but now restricting to non-canonical targets. To identify non-canonical targets we excluded from the analysis all genes whose representative 3′ UTR (as defined in Online Methods) did not contain a match to positions 1-7, 2-8 or 2-7 (in the latter case having also an A opposite position 1 of the miRNA) of the transfected miRNA.

c.7 overlap of functional targets identified by mirza and other methods

Finally, for the 8 miRNAs transfected by Linsley et al. [123], we computed the overlap between functional predictions made by MIRZA and all of the other methods. For each miRNA, and each method m, we determined the number of predictions n_m, that maximizes the estimated number of functional targets for that method. We then separated the targets for both methods into: targets predicted by both, targets predicted only by MIRZA, and targets predicted only by the other method. For each of these 3 subsets of targets we then calculated the fraction f that was downregulated and used this to estimate the total number of functional targets among the set (as described in the online Methods). As a result, we have the estimated number of functional targets predicted by both methods, by MIRZA only, and by the other method only. The results are shown in Suppl. Fig. 47. As in the previous two figures, each panel corresponds to a transfected miRNA, and each bar corresponds to one method. For each bar, the blue section corresponds to the number of functional targets predicted by both MIRZA and the other method, the green section to the number of functional targets predicted by MIRZA only, and the orange bar to the number of functional targets predicted only by the other method.

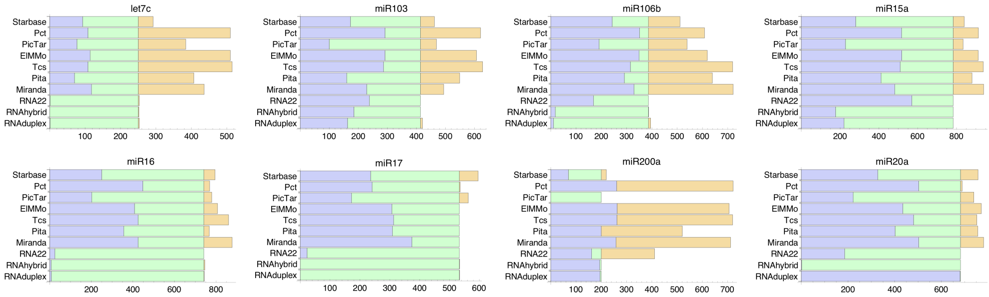


Figure 47: Overlap between MIRZA-predicted functional targets with functional targets predicted by other methods. Each panel corresponds to one transfected miRNA (experimental data from Linsley et al. [123]) and each bar to a prediction method (indicated next to the bar). Colors indicate targets predicted by both methods (blue), targets predicted by MIRZA only (green) and targets predicted by the other method only (orange).

The results show, first of all, that RNA22, RNAduplex, and RNAhybrid hardly ever provide independent functional targets, e.g. functional targets are a combination of some predicted by both and some by MIRZA only. Second, for miR-15a, miR-16, miR-17, and miR-20a MIRZA by itself predicts almost all functional targets, i.e. all other methods contribute only a small number of additional functional sites. In contrast, for miR-200a the functional targets are largely covered by methods such as TargetScan Pct, ElMMo, TargetScan context+, and MIRANDA. For the other examples (let-7c, miR-103, miR-106b) we find that the total set of functional targets is a variable mixture of targets identified by both, targets identified by MIRZA only, and targets identified only by the other method (e.g. TargetScan, ElMMo, or Miranda), suggesting that these methods use complementary information to identify functional targets. Indeed, in order to makes its predictions, MIRZA makes use of Ago2 CLIP data. Although these data were typically obtained in different conditions then those in which the transfection experiments were performed, the fact that a site was detected by CLIP in at least one condition makes the functionality of the site more likely in another condition as well. In contrast, other methods use information, such as conservation, sequence-context, and accessibility, that are not considered by MIRZA. These results thus strongly suggest that a more comprehensive set of miRNA targets could be obtained by combining MIRZA's biophysical model with conservation and context information. We leave such an approach for future work.

Sample	Sites	C	NC	Bulge	miR-124	miR-124	miR-124	miR-124
				@pivot	C	NC	G-bulge	any bulge
							@pivot	@pivot
Mouse Brain A Ago2 130 kDa	5000	2369	2631	98	79	61	5	8
Mouse Brain B Ago2 130 kDa	5000	2362	2638	100	74	70	7	10
Mouse Brain C Ago2 130 kDa	5000	2410	2590	113	105	77	8	11
HeLa miR-124 Tx A Ago2 130 kDa	5000	2005	2995	151	242	480	69	77
HeLa miR-124 Tx B Ago2 130 kDa	5000	1941	3059	127	231	379	45	52
AGO2-HITS-CLIP	5000	3318	1682	56	0	0	0	0
AGO2-PAR-CLIP A	5000	2956	2044	77	0	0	0	0
AGO2-PAR-CLIP B	5000	3445	1555	48	0	0	0	0
AGO2-PAR-CLIP MNase	5000	3574	1426	45	0	0	0	0

Table 1: Relative abundance of various binding modes in CLIP samples (3 Ago2 HITS-CLIP samples from mouse brain and two Ago2 HITS-CLIP samples prepared from miR-124-transfected HeLa cells[26], and the Ago2 CLIP samples used in our study).
C: canonical, NC: non-canonical, pivot is the 6^th position of the miRNA. Bulges occur in the mRNA, between positions 5 and 6 of the miRNA.

BIBLIOGRAPHY

[1] B Alberts, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. *Molecular Biology of the Cell*. Garland Science, 5th edition, 2008. ISBN 9780815341055. URL http://books.google.com/books?id=nMk6PwAACAAJ.

[2] Panagiotis Alexiou, Manolis Maragkakis, Giorgos L Papadopoulos, Martin Reczko, and Artemis G Hatzigeorgiou. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics (Oxford, England)*, 25(23):3049–55, December 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp565. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/23/3049.

[3] Stefan L Ameres, Michael D Horwich, Jui-Hung Hung, Jia Xu, Megha Ghildiyal, Zhiping Weng, and Phillip D Zamore. Target RNA-directed trimming and tailing of small silencing RNAs. *Science (New York, N.Y.)*, 328(5985):1534–9, June 2010. ISSN 1095-9203. doi: 10.1126/science.1187058. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2902985&tool=pmcentrez&rendertype=abstract.

[4] A. Andrus and R.G. Kuimelis. *Base composition analysis of nucleosides using HPLC*, chapter 10. Wiley-interscience, 2001. doi: 10.1002/0471142700.nc1006s01.

[5] Frederick Anokye-Danso, Chinmay M Trivedi, Denise Juhr, Mudit Gupta, Zheng Cui, Ying Tian, Yuzhen Zhang, Wenli Yang, Peter J Gruber, Jonathan A Epstein, and Edward E Morrisey. Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. *Cell stem cell*, 8(4): 376–88, April 2011. ISSN 1875-9777. doi: 10.1016/j.stem.2011.03.001. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3090650&tool=pmcentrez&rendertype=abstract.

[6] U. Atasoy, J. Watson, D. Patel, and J. D. Keene. ELAV protein HuA (HuR) can redistribute between nucleus and cytoplasm and is upregulated during serum stimulation and T cell activation. *J Cell Sci*, 111 (Pt 21):3145–56, 1998.

[7] Sigrid D. Auweter, Florian C. Oberstrass, and Frederic H.-T. Allain. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res.*, 34:4943–4959, 2006.

[8] Paul Babitzke, Carol S Baker, and Tony Romeo. Regulation of translation initiation by RNA binding proteins. *Annual review of microbiology*, 63:27–44, January 2009. ISSN 1545-3251. doi: 10.1146/annurev.micro.091208.073514. URL http://www.ncbi.nlm.nih.gov/pubmed/19385727.

[9] 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.

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

[11] Peter A Beal, Olena Maydanovych, and Subhash Pokharel. The chemistry and biology of RNA editing by adenosine deaminases. *Nucleic acids symposium series (2004)*, pages 83–4, January 2007. ISSN 1746-8272. doi: 10.1093/nass/nrm042. URL http://www.ncbi.nlm.nih.gov/pubmed/18029597.

[12] O. Bembom. *seqLogo: Sequence logos for DNA sequence alignments*, 2008. R package version 1.14.0.

[13] S. Bennett. Solexa ltd. *Pharmacogenomics*, 5:433–438, 2004.

[14] 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.

[15] Doron Betel, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome biology*, 11(8):R90, January 2010. ISSN 1465-6914. doi: 10.1186/gb-2010-11-8-r90. URL http://genomebiology.com/2010/11/8/R90.

[16] S. N. Bhattacharyya, R. Habermacher, U. Martine, E. I. Closs, and W. Filipowicz. Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell*, 125(6):1111–24, 2006.

[17] 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.

[18] 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.

[19] V. Brown, P. Jin, S. Ceman, J. C. Darnell, W. T. O'Donnell, S. A. Tenenbaum, X. Jin, Y. Feng, K. D. Wilkinson, J. D. Keene, R. B. Darnell, and S. T. Warren. Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. *Cell*, 107(4):477–87, 2001.

[20] 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.

[21] J.M. Chambers and T. Hastie. *Statistical models in S*. Wadsworth Brooks Cole computer science series. Wadsworth Brooks/Cole Advanced Books Software, 1992. ISBN 9780534167646. URL http://books.google.com/books?id=uyfvAAAAMAAJ.

[22] 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.

[23] S. W. Chi, J. B. Zang, A. Mele, and R. B. Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–86, 2009.

[24] 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.

[25] S.W. Chi, J.B. Zang, A. Mele, and R.B. Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460: 479–486, 2009.

[26] S.W. Chi, G.J. Hannon, and R.B. Darnell. An alternative mode of microRNA target recognition. *Nat. Struct. Mol. Biol.*, 19:321–327, 2012.

[27] Darren E. Cikaluk, Nasser Tahbaz, Linda C. Hendricks, Gabriel E. DiMattia, Dave Hansen, Dave Pilgrim, and Tom C.

Hobman. GERp95, a Membrane-associated Protein that Belongs to a Family of Proteins Involved in Stem Cell Differentiation. *Mol. Biol. Cell*, 10(10):3357–3372, October 1999. URL http://www.molbiolcell.org/cgi/content/abstract/10/10/3357.

[28] Erica Davis, Florian Caiment, Xavier Tordoir, Jérôme Cavaillé, Anne Ferguson-Smith, Noelle Cockett, Michel Georges, and Carole Charlier. RNAi-mediated allelic trans-interaction at the imprinted Rtl1/Peg11 locus. *Current biology : CB*, 15(8):743–9, April 2005. ISSN 0960-9822. doi: 10.1016/j.cub.2005.02.060. URL http://dx.doi.org/10.1016/j.cub.2005.02.060.

[29] 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.

[30] 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.

[31] C. Dingwall, G. P. Lomonossoff, and R. A. Laskey. High sequence specificity of micrococcal nuclease. *Nucleic Acids Res*, 9(12):2659–73, 1981.

[32] Chuong B Do and Serafim Batzoglou. What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897–9, August 2008. ISSN 1546-1696. doi: 10.1038/nbt1406. URL http://dx.doi.org/10.1038/nbt1406.

[33] 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.

[34] 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.

[35] 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.

[36] C. Ender and G. Meister. Argonaute proteins at a glance. *J Cell Sci*, 123(Pt 11):1819–23, 2010.

[37] T. Eystathioy, E.K. Chan, S.A. Tenenbaum, J.D. Keene, K. Griffith, and M.J. Fritzler. A phosphorylated cytoplasmic autoantigen, GW182, associates with a unique population of human mRNAs within novel cytoplasmic speckles. *Mol. Biol. Cell*, 13:1338–1351, 2002.

[38] X. C. Fan and J. A. Steitz. Overexpression of HuR, a nuclear-cytoplasmic shuttling protein, increases the in vivo stability of ARE-containing mRNAs. *EMBO J*, 17(12):3448–60, 1998.

[39] 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.

[40] 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.

[41] A Fire, S Xu, M K Montgomery, S A Kostas, S E Driver, and C C Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669): 806–11, February 1998. ISSN 0028-0836. doi: 10.1038/35888. URL http://www.ncbi.nlm.nih.gov/pubmed/9486653.

[42] 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.

[43] C. Fraley and A. E. Raftery. MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics, 2006. (revised 2009).

[44] 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.

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

[46] 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.

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

[48] 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(9):e3164, 2008.

[49] David M Garcia, Daehyun Baek, Chanseok Shin, George W Bell, Andrew Grimson, and David P Bartel. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nature structural & molecular biology*, 18(10):1139–46, September 2011. ISSN 1545-9985. doi: 10.1038/nsmb.2115. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3190056&tool=pmcentrez&rendertype=abstract.

[50] Vincenzo Alessandro Gennarino, Marco Sardiello, Raffaella Avellino, Nicola Meola, Vincenza Maselli, Santosh Anand, Luisa Cutillo, Andrea Ballabio, and Sandro Banfi. MicroRNA target prediction by expression analysis of host genes. *Genome research*, 19(3):481–90, March 2009. ISSN 1088-9051. doi: 10.1101/gr.084129.108. URL http://genome.cshlp.org/cgi/content/abstract/gr.084129.108v1.

[51] 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.

[52] André P. Gerber, Daniel Herschlag, and Patrick O. Brown. Extensive Association of Functionally and Cytotopically Related mRNAs with Puf Family RNA-Binding Proteins in Yeast. *PLoS Biology*, 2(3):0342, 2004.

[53] André P Gerber, Stefan Luschnig, Mark A Krasnow, Patrick O Brown, and Daniel Herschlag. Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, 103(12):4487–92, March 2006. ISSN 0027-8424. doi: 10.1073/pnas.0509260103. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1400586&tool=pmcentrez&rendertype=abstract.

[54] 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.

[55] S. Granneman, G. Kudla, E. Petfalski, and D. Tollervey. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc Natl Acad Sci U S A*, 106(24):9613–8, 2009.

[56] 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.

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

[58] Philip A Gregory, Andrew G Bert, Emily L Paterson, Simon C Barry, Anna Tsykin, Gelareh Farshid, Mathew A Vadas, Yeesim Khew-Goodall, and Gregory J Goodall. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nature cell biology*, 10(5):593–601, May 2008. ISSN 1476-4679. doi: 10.1038/ncb1722. URL http://www.ncbi.nlm.nih.gov/pubmed/18376396.

[59] Philip A Gregory, Cameron P Bracken, Andrew G Bert, and Gregory J Goodall. MicroRNAs as regulators of epithelial-mesenchymal transition. *Cell cycle (Georgetown, Tex.)*, 7(20):3112–

8, October 2008. ISSN 1551-4005. URL http://www.ncbi.nlm.nih.gov/pubmed/18927505.

[60] 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.

[61] 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.

[62] 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.

[63] D. Grun, Y.L. Wang, D. Langenberger, K.C. Gunsalus, and N. Rajewsky. microRNA target predictions across seven Drosophila species and comparison to mammalian targets. *PLoS Comp. Biol.*, 1:e13, 2005.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] Andrew J. Hamilton and David C. Baulcombe. A species of small antisense rna in posttranscriptional gene silencing in plants. *Science*, 286(5441):950–952, 1999. doi: 10.1126/science.286.5441.950. URL http://www.sciencemag.org/content/286/5441/950.abstract.

[69] S M Hammond, E Bernstein, D Beach, and G J Hannon. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*, 404(6775):293–6, March 2000. ISSN 0028-0836. doi: 10.1038/35005107. URL http://dx.doi.org/10.1038/35005107.

[70] D.L. Hartl. *Essential Genetics: A Genomics Perspective*. Jones and Bartlett topics in biology. Jones and Bartlett Publishers, 2009. ISBN 9780763773649. URL http://books.google.com/books?id=o5jvsX-9zoIC.

[71] J. Hausser, M. Landthaler, L. Jaskiewicz, D. Gaidatzis, and M. Zavolan. Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome Res.*, 19(11): 2009–2020, 2009.

[72] 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.

[73] 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.

[74] 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.

[75] J. W. Hockensmith, W. L. Kubasek, W. R. Vorachek, and P. H. von Hippel. Laser cross-linking of nucleic acids to proteins. Methodology and first applications to the phage T4 DNA replication system. *J Biol Chem*, 261(8):3512–8, 1986.

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

[77] I. Holmes and R. Durbin. Dynamic programming alignment accuracy. *J. Comput. Biology*, 5:493–504, 1998.

[78] 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.

[79] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409: 860–921, 2001.

[80] D A Jackson, A Pombo, and F Iborra. The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 14(2):242–54, February 2000. ISSN 0892-6638. URL http://www.ncbi.nlm.nih.gov/pubmed/10657981.

[81] Ashwini Jambhekar and Joseph L Derisi. Cis-acting determinants of asymmetric, cytoplasmic RNA transport. *RNA (New York, N.Y.)*, 13(5):625–42, May 2007. ISSN 1355-8382. doi: 10.1261/rna.262607. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1852811&tool=pmcentrez&rendertype=abstract.

[82] M. Kedde, M. J. Strasser, B. Boldajipour, J. A. Oude Vrielink, K. Slanchev, C. le Sage, R. Nagel, P. M. Voorhoeve, J. van Duijse, U. A. Orom, A. H. Lund, A. Perrakis, E. Raz, and R. Agami. RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. *Cell*, 131(7):1273–86, 2007.

[83] M. Kedde, M. van Kouwenhove, W. Zwart, J. A. Oude Vrielink, R. Elkon, and R. Agami. A Pumilio-induced RNA structure switch in p27-3′ UTR controls miR-221 and miR-222 accessibility. *Nat Cell Biol*, 12(10):1014–20, 2010.

[84] J D Keene. RNA regulons: coordination of post-transcriptional events. *Nature Reviews Genetics*, 8:533–543, 2007.

[85] J. D. Keene and S. A. Tenenbaum. Eukaryotic mRNPs may represent posttranscriptional operons. *Mol Cell*, 9(6):1161–7, 2002.

[86] Jack D Keene, Jordan M Komisarow, and Matthew B Friedersdorf. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nature protocols*, 1(1):302–7, January 2006. ISSN 1750-2799. doi: 10.1038/nprot.2006.47. URL http://www.ncbi.nlm.nih.gov/pubmed/17406249.

[87] J.D. Keene and P.J. Lager. Post-transcriptional operons and regulons co-ordinating gene expression. *Chromosome Res.*, 13: 327–337, 2005.

[88] J R Kennerdell and R W Carthew. Use of dsRNA-mediated genetic interference to demonstrate that frizzled and frizzled 2 act in the wingless pathway. *Cell*, 95(7):1017–26, December 1998. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/9875855.

[89] 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.

[90] Mohsen Khorshid, Christoph Rodak, and Mihaela Zavolan. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic acids research*, 39(Database issue):D245–52, January 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq940. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013791&tool=pmcentrez&rendertype=abstract.

[91] Mohsen Khorshid, Jean Hausser, Mihaela Zavolan, and Erik van Nimwegen. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nature methods*, 10(3): 253–5, March 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2341. URL http://dx.doi.org/10.1038/nmeth.2341.

[92] H. H. Kim, Y. Kuwano, S. Srikantan, E. K. Lee, J. L. Martindale, and M. Gorospe. HuR recruits let-7/RISC to repress c-Myc expression. *Genes Dev*, 23(15):1743–8, 2009.

[93] Y Kirino and Z Mourelatos. Site-specific crosslinking of human microRNPs to RNA targets. *RNA*, 14:2254–2259, 2008.

[94] Shivendra Kishore, Sandra Luber, and Mihaela Zavolan. Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Briefings in functional*

genomics, 9(5-6):391–404, December 2010. ISSN 2041-2657. doi: 10.1093/bfgp/elq028. URL http://bfg.oxfordjournals.org/content/early/2010/12/01/bfgp.elq028.full.

[95] Shivendra Kishore, Lukasz Jaskiewicz, Lukas Burger, Jean Hausser, Mohsen Khorshid, and Mihaela Zavolan. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature methods*, 8(7):559–564, January 2011. ISSN 1548-7105. doi: 10.1038/nmeth.1608. URL http://www.ncbi.nlm.nih.gov/pubmed/21572407.

[96] J. König, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, and J. Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, 17(7):909–15, 2010.

[97] J. Krol, V. Busskamp, I. Markiewicz, M. B. Stadler, S. Ribi, J. Richter, J. Duebel, S. Bicker, H. J. Fehling, D. Schubeler, T. G. Oertner, G. Schratt, M. Bibel, B. Roska, and W. Filipowicz. Characterizing light-regulated retinal microRNAs reveals rapid turnover as a common property of neuronal microRNAs. *Cell*, 141(4):618–31, 2010.

[98] 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.

[99] 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.

[100] Mariana Lagos-Quintana, Reinhard Rauhut, Jutta Meyer, Arndt Borkhardt, and Thomas Tuschl. New microRNAs from mouse and human. *RNA (New York, N.Y.)*, 9(2):175–9, February 2003. ISSN 1355-8382. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1370382&tool=pmcentrez&rendertype=abstract.

[101] Eric C Lai, Pavel Tomancak, Robert W Williams, and Gerald M Rubin. Computational identification of Drosophila microRNA genes. *Genome biology*, 4(7):R42, January 2003. ISSN 1465-6914. doi: 10.1186/gb-2003-4-7-r42. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=193629&tool=pmcentrez&rendertype=abstract.

[102] Ashish Lal, Francisco Navarro, Christopher A Maher, Laura E Maliszewski, Nan Yan, Elizabeth O'Day, Dipanjan Chowdhury, Derek M Dykxhoorn, Perry Tsai, Oliver Hofmann, Kevin G Becker, Myriam Gorospe, Winston Hide, and Judy Lieberman. miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements. *Molecular cell*, 35(5):610–25, September 2009. ISSN 1097-4164. doi: 10.1016/j.molcel.2009.08.020. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2757794&tool=pmcentrez&rendertype=abstract.

[103] 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.

[104] 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.

[105] Markus Landthaler, Abdullah Yalcin, and Thomas Tuschl. The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. *Current biology : CB*, 14(23):2162–7, December 2004. ISSN 0960-9822. doi: 10.1016/j.cub.2004.11.001. URL http://dx.doi.org/10.1016/j.cub.2004.11.001.

[106] 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 (New York, N.Y.)*, 14(12):2580–96, December 2008. ISSN 1469-9001. doi: 10.1261/rna.1351608. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2590962&tool=pmcentrez&rendertype=abstract.

[107] B. Langmead, C. Trapnell, M. Pop, and S.L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10:R25, 2009.

[108] N C Lau, Lee P. Lim, E G Weinstein, and David P. Bartel. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294(5543):858–62, 2001. ISSN 0036-8075. doi: 10.1126/science.1065062. URL http://www.ncbi.nlm.nih.gov/pubmed/11679671.

[109] 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.

[110] 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.

[111] Yoontae Lee, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V Narry Kim. MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, 23(20):4051–60, October 2004. ISSN 0261-4189. doi: 10.1038/sj.emboj.7600385. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=524334&tool=pmcentrez&rendertype=abstract.

[112] Young Sik Lee, Kenji Nakahara, John W Pham, Kevin Kim, Zhengying He, Erik J Sontheimer, and Richard W Carthew. Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways. *Cell*, 117(1):69–81, April 2004. ISSN 00928674. doi: 10.1016/S0092-8674(04)00261-2. URL http://dx.doi.org/10.1016/S0092-8674(04)00261-2.

[113] S. K. Leivonen, R. Makela, P. Ostling, P. Kohonen, S. Haapa-Paananen, K. Kleivi, E. Enerly, A. Aakula, K. Hellstrom, N. Sahlberg, V. N. Kristensen, A. L. B?rresen-Dale, P. Saviranta, M. Perala, and O. Kallioniemi. Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines. *Oncogene*, 28(44):3926–3936, Nov 2009.

[114] T. D. Levine, F. Gao, P. H. King, L. G. Andrews, and J. D. Keene. Hel-N1: an autoimmune RNA-binding protein with specificity for 3′ uridylate-rich untranslated regions of growth factor mRNAs. *Mol Cell Biol*, 13(6):3494–504, 1993.

[115] 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.

[116] 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.

[117] H. Li and R. Durbin. Fast and accurate short-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754–1760, 2009.

[118] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26:589–595, 2010.

[119] P Liang and A. Pardee. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257(5072):967–971, August 1992. ISSN 0036-8075. doi: 10.1126/science.1354393. URL http://www.sciencemag.org/content/257/5072/967.abstract.

[120] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell, and R. B. Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–9, 2008.

[121] 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.

[122] 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.

[123] 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.

[124] Ronny Lorenz, Stephan H Bernhart, Christian Hoener zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, November 2011. ISSN 1748-7188. doi: 10.1186/1748-7188-6-26. URL http://www.almob.org/content/6/1/26/abstract.

[125] 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.

[126] 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.

[127] S Madison-Antenucci and S L Hajduk. RNA editing-associated protein 1 is an RNA binding protein with specificity for preedited mRNA. *Molecular cell*, 7(4):879–86, April 2001. ISSN 1097-2765. URL http://www.ncbi.nlm.nih.gov/pubmed/11336710.

[128] K C Martin and A Ephrussi. mRNA Localization: Gene Expression in the Spatial Dimension. *Cell*, 136:719–730, 2009.

[129] Arianne J Matlin, Francis Clark, and Christopher W J Smith. Understanding alternative splicing: towards a cellular code. *Nature reviews. Molecular cell biology*, 6(5):386–98, May 2005. ISSN 1471-0072. doi: 10.1038/nrm1645. URL http://dx.doi.org/10.1038/nrm1645.

[130] V Matys, O V Kel-Margoulis, E Fricke, I Liebich, S Land, A Barre-Dirrie, I Reuter, D Chekmenev, M Krull, K Hornischer, N Voss, P Stegmaier, B Lewicki-Potapov, H Saxel, A E Kel, and E Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(Database issue):D108–10, January 2006. ISSN 1362-4962. doi: 10.1093/nar/gkj143. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1347505&tool=pmcentrez&rendertype=abstract.

[131] 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.

[132] 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.

[133] K M Meisenheimer and T H Koch. Photocross-linking of nucleic acids to associated proteins. *Crit. Rev. Biochem. Mol. Biol.*, 32: 101–140, 1997.

[134] 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.

[135] 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.

[136] M J Moore and N J Proudfoot. Pre-mRNA Processing Reaches Back to Transcription and Ahead to Translation. *Cell*, 136:688–700, 2009.

[137] N. Mukherjee, P.J. Lager, M.B. Friedersdorf, M.A. Thompson, and J.D. Keene. Coordinated posttranscriptional mRNA population dynamics during T-cell activation. *Mol. Syst. Biol.*, 5:288, 2009.

[138] Pierre Neveu, Min Jeong Kye, Shuping Qi, David E Buchholz, Dennis O Clegg, Mustafa Sahin, In-Hyun Park, Kwang-Soo Kim, George Q Daley, Harley I Kornblum, Boris I Shraiman, and Kenneth S Kosik. MicroRNA profiling reveals two distinct p53-related human pluripotent stem cell states. *Cell stem cell*, 7(6):671–81, December 2010. ISSN 1875-9777. doi: 10.1016/j.stem.2010.11.012. URL http://dx.doi.org/10.1016/j.stem.2010.11.012.

[139] 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.

[140] Thomas Ohrt, Wolfgang Staroske, Jörg Mütze, Karin Crell, Markus Landthaler, and Petra Schwille. Fluorescence Cross-Correlation Spectroscopy Reveals Mechanistic Insights into the Effect of 2'-O-Methyl Modified siRNAs in Living Cells. *Biophysical journal*, 100(12):2981–90, June 2011. ISSN 1542-0086. doi: 10.1016/j.bpj.2011.05.005. URL http://www.ncbi.nlm.nih.gov/pubmed/21689532.

[141] Abigail F Olena and James G Patton. Genomic organization of microRNAs. *Journal of cellular physiology*, 222(3):540–5, March 2010. ISSN 1097-4652. doi: 10.1002/jcp.21993. URL http://www.ncbi.nlm.nih.gov/pubmed/20020507.

[142] 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.

[143] Sun-Mi Park, Arti B Gaur, Ernst Lengyel, and Marcus E Peter. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes & development*, 22(7):894–907, April 2008. ISSN 0890-9369. doi: 10.1101/gad.1640608. URL http://genesdev.cshlp.org/cgi/content/abstract/22/7/894.

[144] Simona Pedrotti, Roberta Busà, Claudia Compagnucci, and Claudio Sette. The RNA recognition motif protein RBM11 is a novel tissue-specific splicing regulator. *Nucleic acids research*, October 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr819. URL http://www.ncbi.nlm.nih.gov/pubmed/21984414.

[145] 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.

[146] 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.

[147] D. Ray, H. Kazan, E. T. Chan, L. Pena Castillo, S. Chaudhry, S. Talukder, B. J. Blencowe, Q. Morris, and T. R. Hughes. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol*, 27(7):667–70, 2009.

[148] Marc Rehmsmeier, Peter Steffen, Matthias Hochsmann, and Robert Giegerich. Fast and effective prediction of microRNA/-target duplexes. *RNA*, 10(10):1507–17, October 2004. ISSN 1355-8382. doi: 10.1261/rna.5248604. URL http://www.ncbi.nlm.nih.gov/pubmed/15383676.

[149] 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.

[150] 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.

[151] S. Rudel, A. Flatley, L. Weinmann, E. Kremmer, and G. Meister. A multifunctional human Argonaute2-specific monoclonal antibody. *RNA*, 14(6):1244–53, 2008.

[152] 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.

[153] 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.

[154] R. Schrodinger, E. Schrödinger, and R. Penrose. *What Is Life?: With Mind and Matter and Autobiographical Sketches*. Canto Book. Turtleback Books, 1992. ISBN 9780613920759. URL http://books.google.com/books?id=XLUdAQAACAAJ.

[155] Jennifer A Schwanekamp, Maureen A Sartor, Saikumar Karyala, Danielle Halbleib, Mario Medvedovic, and Craig R Tomlinson. Genome-wide analyses show that nuclear and cytoplasmic RNA levels are differentially affected by dioxin. *Biochimica et biophysica acta*, 1759(8-9):388–402, 2006. ISSN 0006-3002. doi: 10.1016/j.bbaexp.2006.07.005. URL http://dx.doi.org/10.1016/j.bbaexp.2006.07.005.

[156] 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.

[157] Luke A Selth, Chris Gilbert, and Jesper Q Svejstrup. RNA immunoprecipitation to determine RNA-protein associations in vivo. *Cold Spring Harbor protocols*, 2009(6):pdb.prot5234, June 2009. ISSN 1559-6095. doi: 10.1101/pdb.prot5234. URL http://cshprotocols.cshlp.org/cgi/content/abstract/2009/6/pdb.prot5234.

[158] 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.

[159] N G Shoham, T Arad, R Rosin-Abersfeld, P Mashiah, A Gazit, and A Yaniv. Differential display assay and analysis. *BioTechniques*, 20(2):182–4, February 1996. ISSN 0736-6205. URL http://www.ncbi.nlm.nih.gov/pubmed/8825145.

[160] R Siddharthan, E D Siggia, and E van Nimwegen. PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny. *PLoS computational biology*, 1:e67, 2005.

[161] 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.

[162] K.C. Smith and D.H. Meun. Kinetics of the photochemical addition of [35S] cysteine to polynucleotides and nucleic acids. *Biochemistry*, 7:1033–1037, 1968.

[163] G.K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

[164] Nahum Sonenberg and Alan G Hinnebusch. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, 136(4):731–45, February 2009. ISSN 1097-4172. doi: 10.1016/j.cell.2009.01.042. URL http://www.ncbi.nlm.nih.gov/pubmed/19239892.

[165] 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.

[166] L.D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J.E. Stajich, T.W. Harris, A. Arva, and S. Lewis. The generic genome browser: a building block for a model organism system database. *Genome Res.*, 12(10):1599–1610, 2002.

[167] J. Sturtevant. Applications of Differential-Display Reverse Transcription-PCR to Molecular Pathogenesis and Medical Mycology. *Clinical Microbiology Reviews*, 13(3):408–427, July 2000. ISSN 0893-8512. doi: 10.1128/CMR.13.3.408-427.2000. URL http://cmr.asm.org/cgi/content/abstract/13/3/408.

[168] 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.

[169] K. Takahashi and S. Moore. *The Enzymes V*, pages 435–468. Academic Press Inc, New York, 1982.

[170] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–76, August 2006. ISSN 0092-8674. doi: 10.1016/j.cell.2006.07.024. URL http://www.cell.com/fulltext/S0092-8674(06)00976-7.

[171] 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.

[172] 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.

[173] 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.

[174] Elizabeth J Thatcher, Jordan Bond, Ima Paydar, and James G Patton. Genomic organization of zebrafish microRNAs. *BMC genomics*, 9:253, January 2008. ISSN 1471-2164. doi: 10.1186/1471-2164-9-253. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2427041&tool=pmcentrez&rendertype=abstract.

[175] 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.

[176] 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(5648):1212–5, 2003.

[177] J. Ule, K. Jensen, A. Mele, and R. B. Darnell. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods*, 37(4):376–86, 2005.

[178] 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.

[179] J.M. Valcarcel, J. Izquierdo. Two isoforms of the T-cell intracellular antigen 1 (TIA-1) splicing factor display distinct splicing

regulation activities. Control of TIA-1 isoform ratio by TIA-1-related protein. *J. Biol. Chem.*, 282:19410–19417, 2007.

[180] 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.

[181] 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.

[182] 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.

[183] 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.

[184] X. Wang and T. M. Tanaka Hall. Structural basis for recognition of AU-rich element RNA by the HuD protein. *Nat Struct Biol*, 8 (2):141–5, 2001.

[185] 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.

[186] 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.

[187] 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.

[188] Z. Wang, J. Tollervey, M. Briese, D. Turner, and J. Ule. CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo. *Methods*, 48(3): 287–93, 2009.

[189] R.F. Weaver. *Molecular Biology*. McGraw-Hill, 2007. ISBN 9780073319940. URL http://books.google.com/books?id=M-Q9PgAACAAJ.

[190] 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.

[191] 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.

[192] T.D. Wu and S. Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26: 873–881, 2010.

[193] T.D. Wu and C.K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21:1859–1875, 2005.

[194] 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.

[195] 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.

[196] J.H Yang, J.H. Li, P. Shao, H. Zhou, Y.Q. Chen, and L.H. Qu. starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucl. Acids Res.*, 39:D202–D209, 2010.

[197] Soraya Yekta, I-Hung Shih, and David P Bartel. MicroRNA-directed cleavage of HOXB8 mRNA. *Science (New York, N.Y.)*, 304(5670):594–6, April 2004. ISSN 1095-9203. doi: 10.1126/science.1097434. URL http://www.sciencemag.org/content/304/5670/594.abstract.

[198] 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.

[199] J K Yisraeli. VICKZ proteins: a multi-talented family of regulatory RNA-binding proteins. Biol. *Cell*, 97:87–96, 2005.

[200] S. Zheng, T.A. Robertson, and G. Varani. A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. *FEBS JOURNAL*, 274(24):6378–6391, 2007.

[201] 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.