Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP

Markus Hafner,[1,5] Markus Landthaler,[1,4,5] Lukas Burger,[2] Mohsen Khorshid,[2] Jean Hausser,[2] Philipp Berninger,[2] Andrea Rothballer,[1] Manuel Ascano, Jr.,[1] Anna-Carina Jungkamp,[1,4] Mathias Munschauer,[1] Alexander Ulrich,[1] Greg S. Wardle,[1] Scott Dewell,[3] Mihaela Zavolan,[2,*] and Thomas Tuschl[1,*]

[1]Howard Hughes Medical Institute, Laboratory of RNA Molecular Biology, The Rockefeller University, 1230 York Avenue, Box 186, New York, NY 10065, USA

[2]Biozentrum der Universität Basel and Swiss Institute of Bioinformatics (SIB), Klingelbergstr. 50-70, CH-4056 Basel, Switzerland

[3]Genomics Resource Center, The Rockefeller University, 1230 York Avenue, Box 241, New York, NY 10065, USA

[4]Present address: Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine, 13125 Berlin, Germany

[5]These authors contributed equally to this work

*Correspondence: mihaela.zavolan@unibas.ch (M.Z.), ttuschl@rockefeller.edu (T.T.)

DOI 10.1016/j.cell.2010.03.009

SUMMARY

RNA transcripts are subject to posttranscriptional 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.

INTRODUCTION

Gene expression in eukaryotes is extensively controlled at the posttranscriptional level by RNA-binding proteins (RBPs) and ribonucleoprotein complexes (RNPs) modulating the maturation, stability, transport, editing and translation of RNA transcripts (Martin and Ephrussi, 2009; Moore and Proudfoot, 2009; Sonenberg and Hinnebusch, 2009). Vertebrate genomes encode several hundred RBPs (McKee et al., 2005), each containing one or more domains able to specifically recognize target transcripts. Furthermore, hundreds of microRNAs (miRNAs) bound by Argonaute (AGO/EIF2C) proteins mediate destabilization and/or inhibition of translation of partially complementary target mRNAs (Bartel, 2009). 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 (Keene, 2007).

A combination of genetic, biochemical and computational approaches are typically applied to identify RNA-RBP or RNA-RNP interactions. Microarray profiling of RNAs associated with immunopurified RBPs (RIP-Chip) (Tenenbaum et al., 2000) 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 (Lopez de Silanes et al., 2004) or for Pumilio (Gerber et al., 2006).

More direct RBP target site information is obtained by combining in vivo UV crosslinking (Greenberg, 1979; Wagenmakers et al., 1980) with immunoprecipitation (Dreyfuss et al., 1984; Mayrand et al., 1981) followed by the isolation of crosslinked RNA segments and cDNA sequencing (CLIP) (Ule et al., 2003). CLIP was used to identify targets of the splicing regulators NOVA1 (Licatalosi et al., 2008), FOX2 (Yeo et al., 2009) and SFRS1 (Sanford et al., 2009) as well as U3 snoRNA and pre-rRNA (Granneman et al., 2009), pri-miRNA targets for HNRNPA1 (Guil and Caceres, 2007), EIF2C2/AGO2 protein binding sites (Chi et al., 2009) and ALG-1 target sites in *C. elegans* (Zisoulis et al., 2010). 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 noncrosslinked 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 (*P*hoto*a*ctivatable-*R*ibonucleoside-Enhanced *C*ross*l*inking

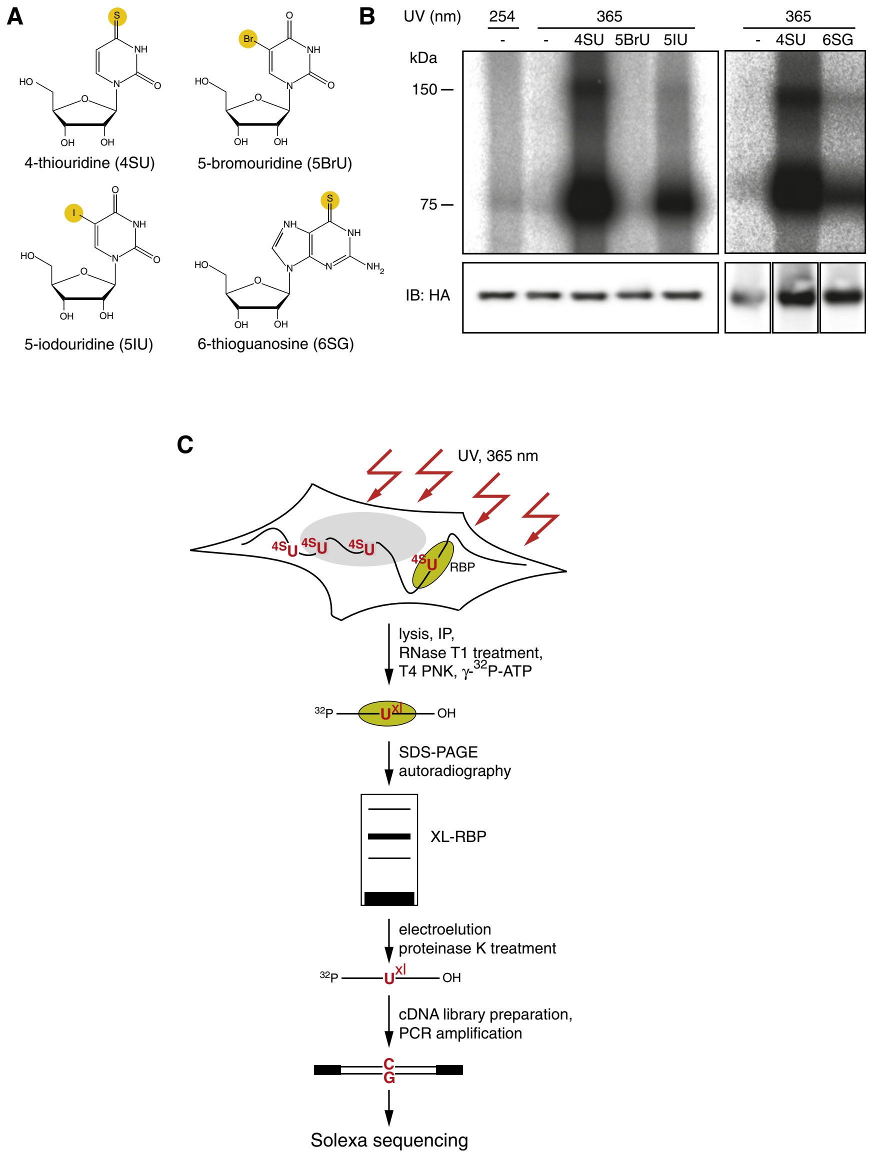


Figure 1. PAR-CLIP Methodology

(A) Structure of photoactivatable nucleosides.

(B) Phosphorimages of SDS-gels that resolved 5′-^{32}P-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.

and *I*mmuno*p*recipitation). 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 posttranscriptional regulation of cellular systems.

RESULTS

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 (Kirino and Mourelatos, 2008; Meisenheimer and Koch, 1997). Several of these photoactivatable nucleosides are readily taken up by cells without apparent toxicity and have been used for in vivo crosslinking (Favre et al., 1986). 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 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 hr by mRNA microarrays. The mRNA profiles of 4SU or 6SG treated cells were very similar to those of untreated cells (Table S1 available online), suggesting that the conditions for endogenous labeling of transcripts were not toxic.

To guide the development of bioinformatic methods for identification of binding sites, we first studied human Pumilio 2 (PUM2), a member of the Puf-protein family (Figure 2A) known for its highly sequence-specific RNA binding (Wang et al., 2002).

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 (Hafner et al., 2008). The sequence reads were aligned against the human genome and EST databases. Reads mapping uniquely to the genome

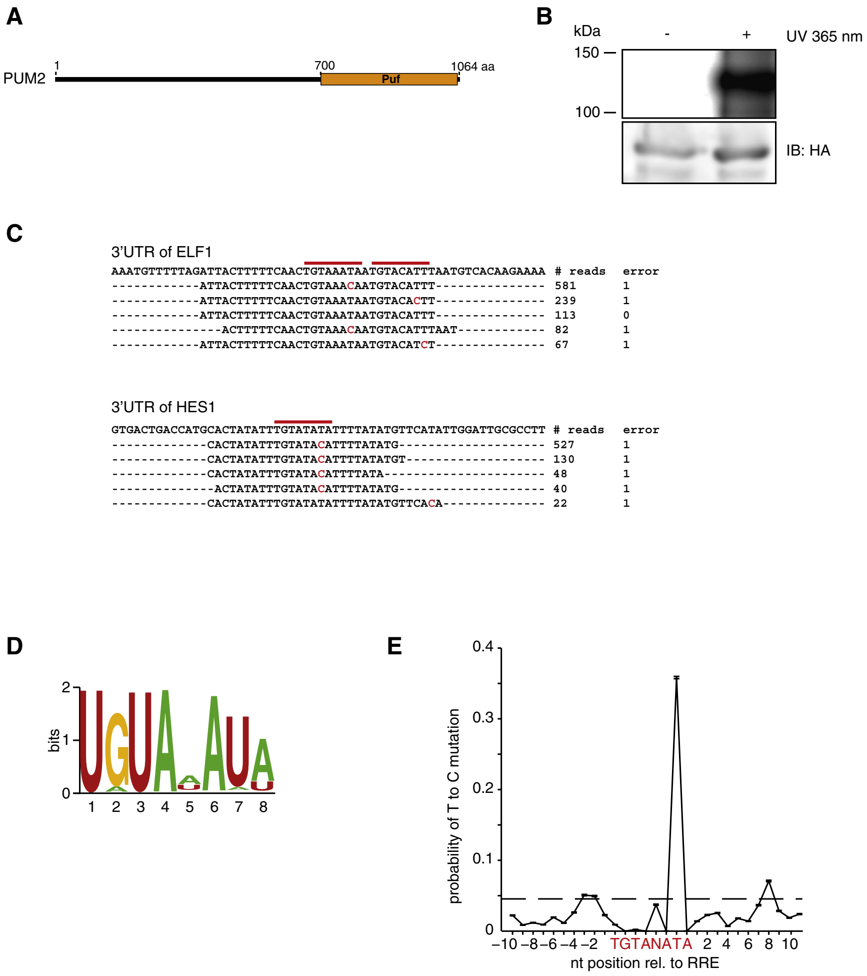


Figure 2. RNA Recognition by PUM2 Protein

(A) Domain structure of PUM2 protein.

(B) Phosphorimage of SDS-gel of radiolabeled FLAG/HA-PUM2-RNA complexes from nonirradiated 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 S1.

with up to one mismatch, insertion or deletion were used to build clusters of sequence reads (Figure 2C, Extended Experimental Procedures, and Table S2). We obtained 7523 clusters originating from about 3000 unique transcripts, 93% of which were found within the 3′ untranslated region (UTR) (Figure S1) in agreement with previous studies (Wickens et al., 2002). All sequence clusters with mapping and annotation information are available online (http://www.mirz.unibas.ch/restricted/clipdata/RESULTS/index.html).

PhyloGibbs analysis (Siddharthan et al., 2005) of the top 100 most abundantly sequenced clusters (Table S3), as expected, yielded the PUM2 RRE, UGUANAUA (Galgano et al., 2008) (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 S1). Ranking of sequence read clusters according to the frequency of T to C mutation further enriched for the PUM2 RRE (Figure S1) 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.

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 and 3B). The RRE ACUAAY was determined by SELEX (Galarneau and Richard, 2005), 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 (Chenard and Richard, 2008).

PhyloGibbs analysis of the 100 most abundantly sequenced clusters (Table S3) yielded the RRE AYUAAY (Figures 3C and 3D), similar to a motif identified by SELEX (Galarneau and Richard, 2005). We found approximately 6000 clusters mapping to 2500 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 S2).

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

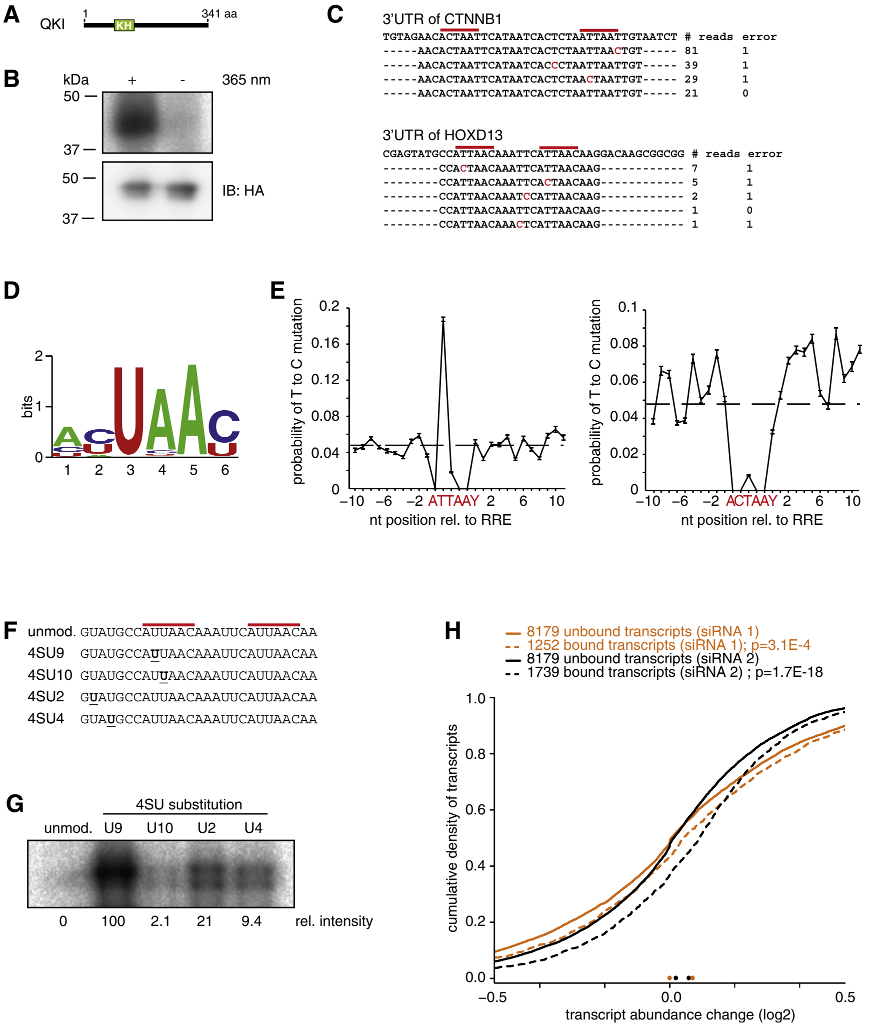


Figure 3. RNA Recognition by QKI Protein

(A) Domain structure of QKI protein.

(B) Phosphorimage of SDS-gel resolving radiolabeled RNA crosslinked to FLAG/HA-QKI IPs from nonirradiated 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. Two distinct siRNA duplexes (1, orange traces and 2, black traces) were used for QKI knockdown and changes in transcript stability relative to mock transfection were inferred from 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. The p-values obtained in the Wilcoxon rank-sum test comparing the changes in targeted and nontargeted transcripts are indicated.

See also Figure S2.

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.

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 and 3G). 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 noncrosslinked 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 nonirradiated 4SU-containing oligoribonucleotides varied between 10 and 20%, and increased to 50% to 80% upon crosslinking (Table S1). The variation of the degree of T to C changes in the crosslinked samples is most likely determined by background of noncrosslinked 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 S1). 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 noncrosslinked 4SU-labeled site, we estimated that the average T to C conversion rate of 40-nt sequence reads derived from background noncrosslinked sequences will

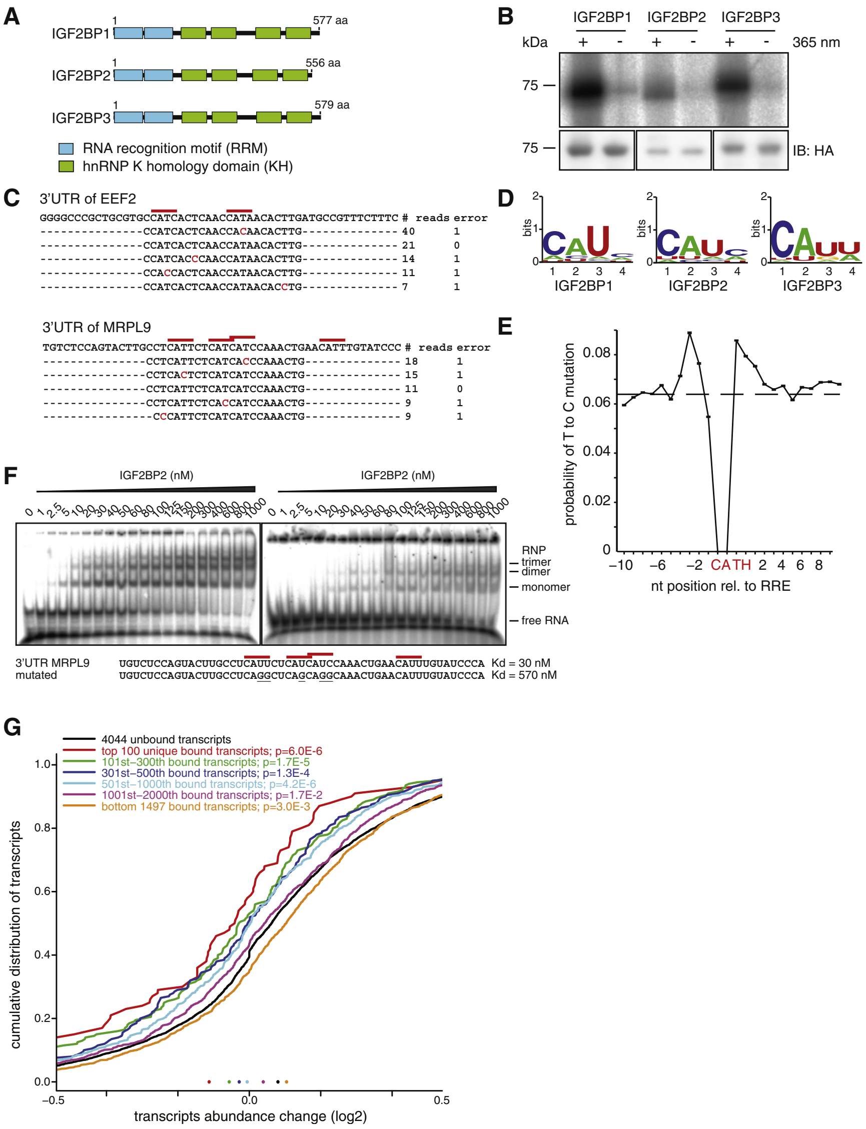


Figure 4. RNA Recognition by the IGF2BP Protein Family

(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 (K_d) are indicated.

(G) Destabilization of IGF2BP-bound transcripts upon siRNA knockdown. A cocktail of three siRNA duplexes targeting IGF2BP1, 2, and 3 was used, as well as a mock transfection and changes in transcript stability were monitored by microarray analysis. Distributions of transcript level changes for IGF2BP1-3 PAR-CLIP target transcripts versus nontargeted transcripts are shown. IGF2BP1-3 target sequences were ranked and divided into bins. The p-values indicate the significance of the difference between the changes of target versus nontarget transcripts, as given by the Wilcoxon rank-sum test and are corrected for multiple testing.

See also Figure S3 and Figure S4.

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 (Chenard and Richard, 2008).

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 and 4B), a family of highly conserved proteins that play a role in cell polarity and cell proliferation (Yisraeli, 2005). These proteins are predominantly expressed in the embryo and regulate mRNA stability, transport and translation. They are re-expressed in various cancers (Boyerinas et al., 2008; Dimitriadis et al., 2007) and IGF2BP2 has been associated with type-2 diabetes (Diabetes Genetics Initiative of Broad Institute of Harvard and MIT et al., 2007). 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 noncoding RNA targets being known so far. A precise definition of the RREs is missing (Yisraeli, 2005).

The three IGF2BPs recognized a highly similar set of target transcripts (Table S1), suggesting similar and redundant functions. PhyloGibbs analysis of the clusters derived from mRNAs (Figure 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 S3). 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 S3). 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 S13). 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 (Yisraeli, 2005), 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 (Licatalosi et al., 2008; Ule et al., 2003), 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 S4).

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”) (Bartel, 2009). Target sites were thought to be predominantly located in the 3′UTRs of mRNAs, and computational miRNA target prediction methods frequently resort to identification of evolutionarily conserved sites that are located in 3′UTRs and are complementary to miRNA seed regions (Bartel, 2009; Rajewsky, 2006).

We isolated mRNA fragments bound by miRNPs from HEK293 cell lines stably expressing FLAG/HA-tagged AGO or TNRC6 family proteins (Landthaler et al., 2008). 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 4000 clusters that overlapped, supporting our earlier observation that AGO1-4 bound similar sets of transcripts (Landthaler et al., 2008). 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 4647 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 (Landthaler et al., 2008). We therefore combined all sequence reads from all TNRC6 experiments, yielding 1865 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 S5).

Comparison of miRNA Profiles from AGO PAR-CLIP to Noncrosslinked miRNA profiles

To relate the potential miRNA-target-site–containing CCRs to the endogenously expressed miRNAs, we determined the miRNA profiles from total RNA isolated from HEK293 cells, and miRNAs isolated from noncrosslinked AGO1-4 IPs by Solexa sequencing (Hafner et al., 2008), 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 noncrosslinked cells correlated well (Figure 5C and Table S5) supporting our observation that AGO1-4 bind the same targets (Landthaler et al., 2008). 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 noncrosslinked libraries showed a good correlation (Spearman correlation coefficient of 0.56, Figure 5C and Figure S5A).

Importantly, in the AGO PAR-CLIP library, the majority of miRNA sequence reads derived from prototypical miRNAs (Landgraf et al., 2007) 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 (Wang et al., 2008). Five of the 100 most abundant miRNAs in HEK293 cells lack uridines at position 8–13, yet only 2 of those miRNAs,

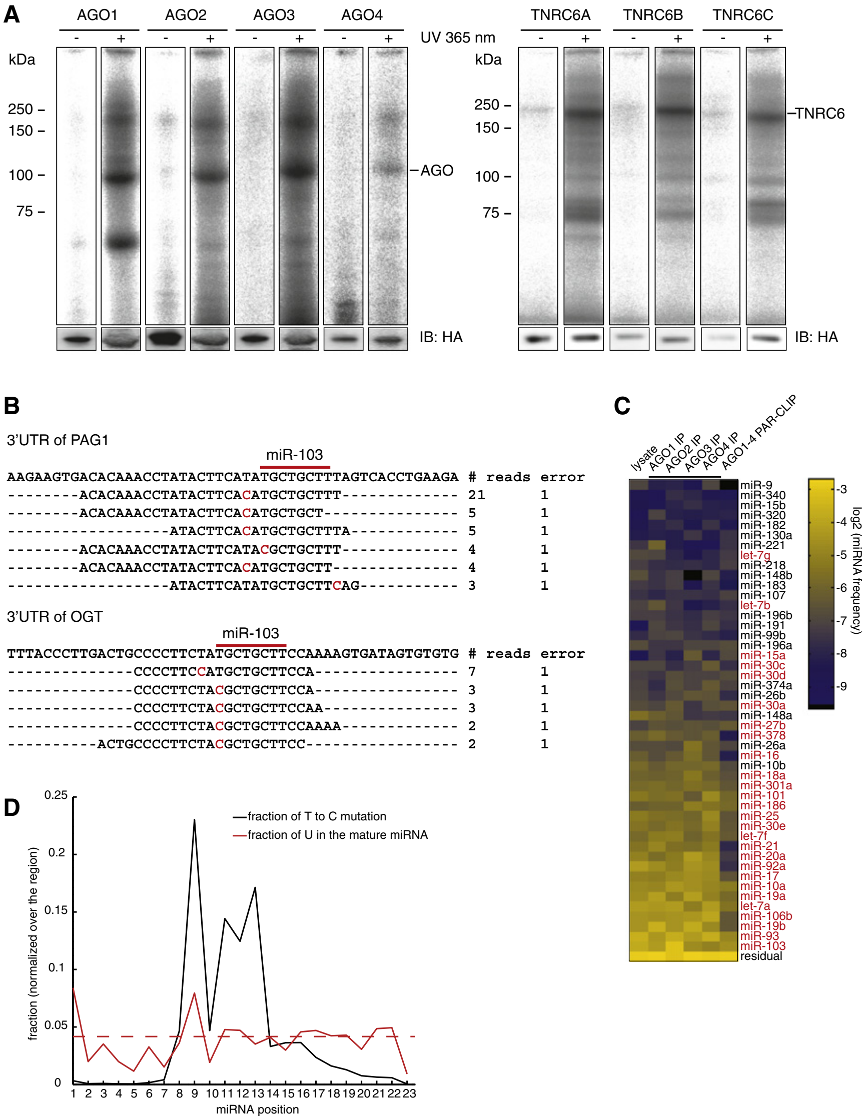


Figure 5. AGO Protein Family and TNRC6 Family PAR-CLIP

(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, noncrosslinked 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 S5.

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 S5).

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 center of the AGO-miRNA-target-RNA ternary complex, where the target RNA is proximal to the Piwi/RNase H domain of the AGO protein (Wang et al., 2008). 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 (Wang et al., 2008, 2009). 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 (Grimson et al., 2007; Hausser et al., 2009).

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.

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

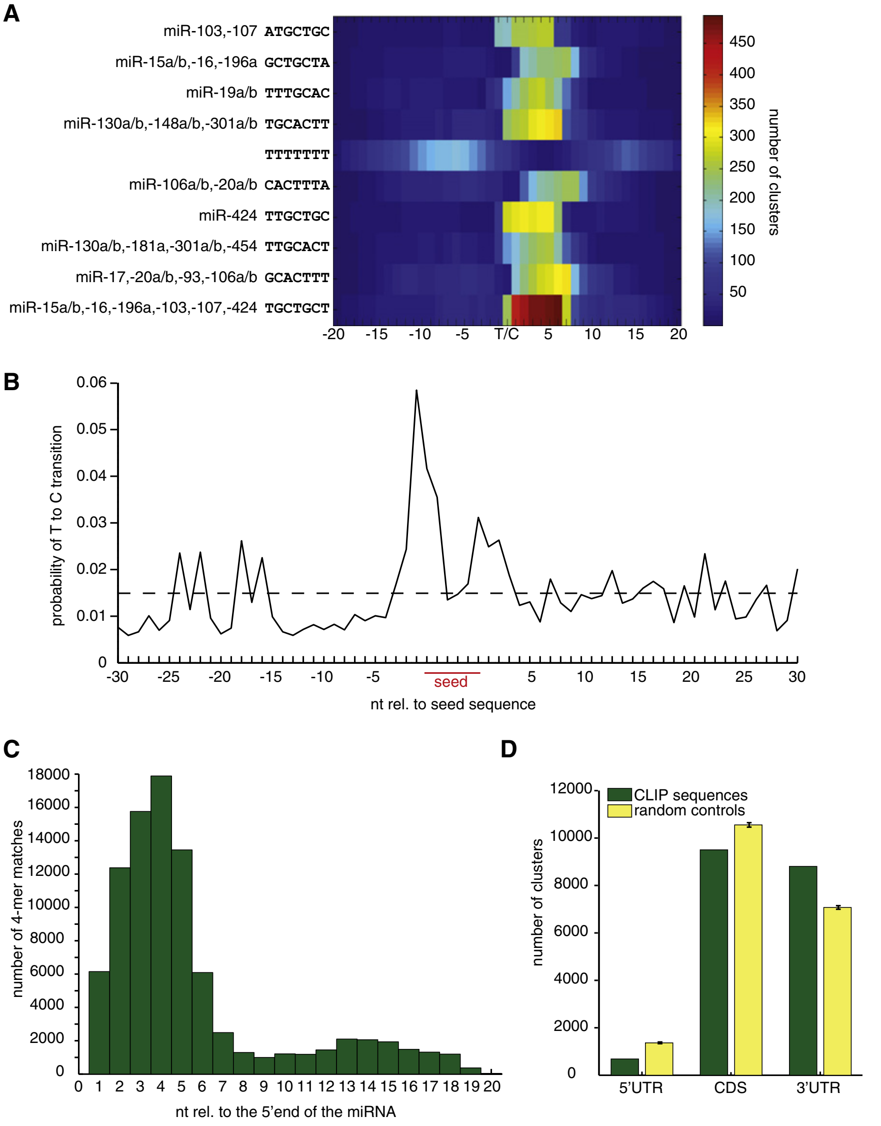


Figure 6. AGO PAR-CLIP Identifies miRNA Seed-Complementary Sequences in HEK293 Cells

(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 S6.

50% of the CCRs contain 6-mers corresponding to one of the top 100 expressed miRNAs (Figure S5), 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 (Gaidatzis et al., 2007; Lewis et al., 2005) (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 S6B).

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 S6C), which was expected from previous bioinformatic analyses of functional miRNA-binding sites.

Noncanonical 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 (Wang et al., 2008). 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 (Grimson et al., 2007), we also detected

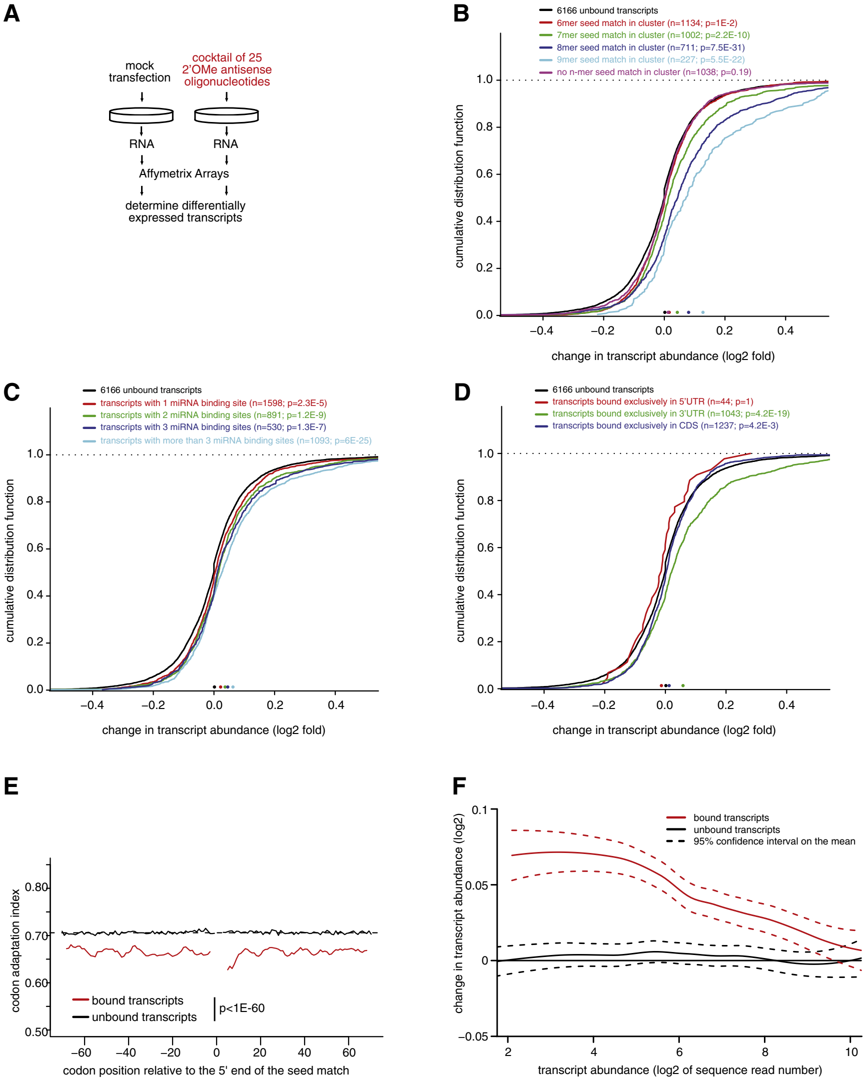


Figure 7. Relationship between Various Features of miRNA/Target RNA Interactions and mRNA Stability

(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. The stability change for transcripts harboring CCRs without identifiable miRNA seed-complementary regions is also shown. The p values indicate the significance of the difference between the transcript level changes of transcripts containing CCRs versus transcripts without CCRs, as given by the Wilcoxon rank-sum test and are corrected for multiple testing.

(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. The red and the black lines indicate the CAI for seed-complementary sequence containing transcripts bound and not bound by AGO proteins determined by AGO PAR-CLIP.

(F) LOESS regression of total transcript abundance in HEK293 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 S7.

a weak signal for contiguous 4-nt long matches to positions 13–15 of the miRNA (Figure 6C).

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 (Easow et al., 2007; Lewis et al., 2005). However, miRNAs are believed to predominantly act on 3′UTRs (Bartel, 2009), with relatively few reports providing experimental evidence for miRNA-binding to individual 5′UTRs or CDS (Easow et al., 2007; Forman et al., 2008; Lytle et al., 2007; Orom et al., 2008; Tay et al., 2008).

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 (Grimson et al., 2007), 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 S6). Both of these findings are in agreement with previous observations (Grimson et al., 2007).

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) (Sharp and Li, 1987) around crosslinked seed matches, and found that the sequence environment of crosslinked seed matches differed from that of noncrosslinked 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 miRNA-target-RNA interactions as ribosomes are stalled (Gu et al., 2009). 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 (Stark et al., 2005), 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 (Baek et al., 2008; Lim et al., 2005; Selbach et al., 2008). We transfected miRNA duplexes corresponding to the deeply conserved miR-7 and miR-124 into FLAG/HA-AGO2 cells, performed PAR-CLIP (Figure S7), 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 S7).

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 (Gaidatzis et al., 2007), the mRNA stability scores by TargetScan context score (Grimson et al., 2007), and sequence composition and structure measures for the regions around the miRNA seed complementary sites. The feature that distinguished most crosslinked from noncrosslinked seed matches was a 25% lower free energy required to resolve local secondary structure involving the miRNA-binding region (Figure S7), 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 noncrosslinked 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 nonrandom 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 S7). The location of CCRs in the CDS was biased toward the stop codon for the transfected miR-7 and 124, but not for the endogenous miRNAs (Figure S7).

Finally, we wanted to examine how miRNA targets defined by PAR-CLIP compared in regulation of target mRNA stability to those predicted by ElMMo (Gaidatzis et al., 2007), TargetScan context score (Grimson et al., 2007), TargetScan Pct (Friedman et al., 2009) and PicTar (Lall et al., 2006). 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 S7).

DISCUSSION

Maturation, localization, decay and translational regulation of mRNAs involve formation of complexes of RBPs and RNPs with their RNA targets (Martin and Ephrussi, 2009; Moore and Proudfoot, 2009). 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 (Lunde et al., 2007). Furthermore hundreds of miRNAs function together with AGO and TNRC6 proteins to destabilize target mRNAs and/or repress their translation (Bartel, 2009). Collectively, these factors and their presumably combinatorial action constitute the code for posttranscriptional gene regulation. Here we describe an approach to directly identify transcriptome-wide mRNA-binding sites of regulatory RBPs and RNPs in live cells.

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 (Chi et al., 2009; Licatalosi et al., 2008; Ule et al., 2003;

Zisoulis et al., 2010), except for recent work by Grannemann et al. on the U3 snoRNA that showed an increase of deletions at the RBP binding site (Granneman et al., 2009). 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 S3).

From an experimental perspective, it is important to note that crosslinked RNA segments, irrespective of the methods of isolation, are always contaminated with noncrosslinked 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 noncrosslinked noise. We now provide a method that accomplishes this critical task.

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.

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 (Gu et al., 2009).

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 noncanonical miRNA-binding modes. We were able to explain the majority of PAR-CLIP binding sites by conventional miRNA-mRNA seed-pairing interactions (Grimson et al., 2007), 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 (Vella et al., 2004) and those recently observed in biochemical and structural studies of *T. thermophilus* Ago protein (Wang et al., 2008; Wang et al., 2009).

The mRNA Ribonucleoprotein 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 posttranscriptional gene regulation.

EXPERIMENTAL PROCEDURES

PAR-CLIP

Human embryonic kidney (HEK) 293 cells stably expressing FLAG/HA-tagged IGF2BP1-3, QKI, PUM2, AGO1-4, and TNRC6A-C (Landthaler et al., 2008) 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 (Hafner et al., 2008) 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 Extended Experimental Procedures. For a video presenting the procedure please visit http://www.jove.com/index/Details.stp?ID=2034.

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.

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 Supplemental Information.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, and seven tables and can be found with this article online at doi:10.1016/j.cell.2010.03.009.

ACKNOWLEDGMENTS

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

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

Received: June 23, 2009
Revised: January 11, 2010
Accepted: February 27, 2010
Published: April 1, 2010

REFERENCES

Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64–71.

Bartel, D.P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. Cell *136*, 215–233.

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

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

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

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, L.U.a.N.I.o.B.R., Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I.W., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N., et al. (2007). Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. Science *316*, 1331–1336.

Dimitriadis, E., Trangas, T., Milatos, S., Foukas, P.G., Gioulbasanis, I., Courtis, N., Nielsen, F.C., Pandis, N., Dafni, U., Bardi, G., et al. (2007). Expression of oncofetal RNA-binding protein CRD-BP/IMP1 predicts clinical outcome in colon cancer. Int. J. Cancer *121*, 486–494.

Dreyfuss, G., Choi, Y.D., and Adam, S.A. (1984). Characterization of heterogeneous nuclear RNA-protein complexes in vivo with monoclonal antibodies. Mol. Cell. Biol. *4*, 1104–1114.

Easow, G., Teleman, A.A., and Cohen, S.M. (2007). Isolation of microRNA targets by miRNP immunopurification. RNA *13*, 1198–1204.

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

Forman, J.J., Legesse-Miller, A., and Coller, H.A. (2008). A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proc. Natl. Acad. Sci. USA *105*, 14879–14884.

Friedman, R.C., Farh, K.K.-H., Burge, C.B., and Bartel, D. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. *19*, 92–105.

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

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

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

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

Granneman, S., Kudla, G., Petfalski, E., and Tollervey, D. (2009). Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. Proc. Nat. Acad. Sci. *106*, 9613–9618.

Greenberg, J.R. (1979). Ultraviolet light-induced crosslinking of mRNA to proteins. Nucleic Acids Res. *6*, 715–732.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol. Cell *27*, 91–105.

Gu, S., Jin, L., Zhang, F., Sarnow, P., and Kay, M.A. (2009). Biological basis for restriction of microRNA targets to the 3′ untranslated region in mammalian mRNAs. Nat. Struct. Mol. Biol. *16*, 144–150.

Guil, S., and Caceres, J.F. (2007). The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. Nat. Struct. Mol. Biol. *14*, 591.

Hafner, M., Landgraf, P., Ludwig, J., Rice, A., Ojo, T., Lin, C., Holoch, D., Lim, C., and Tuschl, T. (2008). Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods *44*, 3–12.

Hausser, J., Landthaler, M., Jaskiewicz, L., Gaidatzis, D., and Zavolan, M. (2009). 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*, 2009–2020.

Keene, J.D. (2007). RNA regulons: coordination of post-transcriptional events. Nat. Rev. Genet. *8*, 533–543.

Kirino, Y., and Mourelatos, Z. (2008). Site-specific crosslinking of human microRNPs to RNA targets. RNA *14*, 2254–2259.

Lall, S., Grun, D., Krek, A., Chen, K., Wang, Y.-L., Dewey, C.N., Sood, P., Colombo, T., Bray, N., MacMenamin, P., et al. (2006). A Genome-Wide Map of Conserved MicroRNA Targets in C. elegans. Curr. Biol. *16*, 460–471.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M., et al. (2007). A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. Cell *129*, 1401–1414.

Landthaler, M., Gaidatzis, D., Rothballer, A., Chen, P.Y., Soll, S.J., Dinic, L., Ojo, T., Hafner, M., Zavolan, M., and Tuschl, T. (2008). Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. RNA *14*, 2580–2596.

Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15–20.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature *433*, 769–773.

Lopez de Silanes, I., Zhan, M., Lal, A., Yang, X., and Gorospe, M. (2004). Identification of a target RNA motif for RNA-binding protein HuR. Proc. Natl. Acad. Sci. USA *101*, 2987–2992.

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat. Rev. Mol. Cell Biol. *8*, 479–490.

Lytle, J.R., Yario, T.A., and Steitz, J.A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. Proc. Natl. Acad. Sci. USA *104*, 9667–9672.

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

Mayrand, S., Setyono, B., Greenberg, J.R., and Pederson, T. (1981). Structure of nuclear ribonucleoprotein: identification of proteins in contact with poly(A)+ heterogeneous nuclear RNA in living HeLa cells. J. Cell Biol. *90*, 380–384.

McKee, A.E., Minet, E., Stern, C., Riahi, S., Stiles, C.D., and Silver, P.A. (2005). 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.

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

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

Orom, U.A., Nielsen, F.C., and Lund, A.H. (2008). MicroRNA-10a Binds the 5′UTR of Ribosomal Protein mRNAs and Enhances Their Translation. Mol. Cell *30*, 460–471.

Rajewsky, N. (2006). microRNA target predictions in animals. Nat. Genet. *38*, S8–S13.

Sanford, J.R., Wang, X., Mort, M., Vanduyn, N., Cooper, D.N., Mooney, S.D., Edenberg, H.J., and Liu, Y. (2009). Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome Res. *19*, 381–394.

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58–63.

Sharp, P.M., and Li, W.H. (1987). The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. *15*, 1281–1295.

Siddharthan, R., Siggia, E.D., and van Nimwegen, E. (2005). PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny. PLoS Comp. Biol. *1*, e67.

Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. Cell *136*, 731–745.

Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. (2005). Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3′UTR Evolution. Cell *123*, 1133–1146.

Tay, Y., Zhang, J., Thomson, A.M., Lim, B., and Rigoutsos, I. (2008). MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature *455*, 1124–1128.

Tenenbaum, S.A., Carson, C.C., Lager, P.J., and Keene, J.D. (2000). Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. Proc. Natl. Acad. Sci. USA *97*, 14085–14090.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212–1215.

Vella, M.C., Choi, E.Y., Lin, S.Y., Reinert, K., and Slack, F.J. (2004). The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3′UTR. Genes Dev. *18*, 132–137.

Wagenmakers, A.J., Reinders, R.J., and van Venrooij, W.J. (1980). Cross-linking of mRNA to proteins by irradiation of intact cells with ultraviolet light. Eur. J. Biochem. *112*, 323–330.

Wang, X., McLachlan, J., Zamore, P.D., and Hall, T.M.T. (2002). Modular Recognition of RNA by a Human Pumilio-Homology Domain. Cell *110*, 501–512.

Wang, Y., Juranek, S., Li, H., Sheng, G., Tuschl, T., and Patel, D.J. (2008). Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. Nature *456*, 921–926.

Wang, Y., Juranek, S., Li, H., Sheng, G., Wardle, G.S., Tuschl, T., and Patel, D.J. (2009). Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. Nature *461*, 754–761.

Wickens, M., Bernstein, D.S., Kimble, J., and Parker, R. (2002). A PUF family portrait: 3′UTR regulation as a way of life. Trends Genet. *18*, 150–157.

Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.-D., and Gage, F.H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat. Struct. Mol. Biol. *16*, 130–137.

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

Zisoulis, D.G., Lovci, M.T., Wilbert, M.L., Hutt, K.R., Liang, T.Y., Pasquinelli, A.E., and Yeo, G.W. (2010). Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. Nat. Struct. Mol. Biol. *17*, 173–179.

Supplemental Information

EXTENDED 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 (Dharmacon):

anti-let-7a: AACUAUACAACCUACUACCUCA-NH_2; -NH_2 indicates C6 aminolinker (Dharmacon).

anti-miR-10a: CACAAAUUCGGAUCUACAGGGUA-NH_2;
anti-miR-15a: CGCCAAUAUUUACGUGCUGCUA;
anti-miR-15b: CACAAACCAUUAUGUGCUGCUA;
anti-miR-16: UGUAAACCAUGAUGUGCUGCUA;
anti-miR-17-5p: CUACCUGCACUGUAAGCACUUUG;
anti-miR-18a: CUAUCUGCACUAGAUGCACCUUA-NH_2;
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-NH_2;
anti-miR-30b/c: GAGUGUAGGAUGUUUACA-NH_2;
anti-miR-92b: ACAGGCCGGGACAAGUGCAAUA;
anti-miR-93: CUACCUGCACGAACAGCACUUUG;
anti-miR-101: UUCAGUUAUCACAGUACUGUA;
anti-miR-103: UCAUAGCCCUGUACAAUGCUGCU;
anti-miR-106b: AUCUGCACUGUCAGCACUUUA-NH_2;
anti-miR-186: AGCCCAAAAGGAGAAUUCUUUG;
anti-miR-301: GCUUUGACAAUACUAUUGCACUG;
anti-miR-378: CCUUCUGACUCCAAGUCCAGU;
miR-7/miR-7* duplex, UGGAAGACUAGUGAUUUUGUUGU, CAACAAAUCACAGUCUGCCAUA;
miR-124/miR-124* duplex, 5′-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 (Invitrogen) using GATEWAY LR recombinase (Invitrogen) according to manufacturer's protocol to allow for doxycycline-inducible expression of stably transfected FLAG/HA-tagged protein in Flp-In T-REx HEK293 cells (Invitrogen) from the TO/CMV promoter. pENTR4 QKI and pENTR4 PUM2 were recombined into pFRT/FLAG/HA-DEST for constitutive expression in Flp-In T-REx HEK293 cells.

Plasmids for bacterial expression of N-terminally His_6-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 His_6-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 His_6-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 hr 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 12,000 × g 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 hr. 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. His_6-IGF2BP1, −2, and −3 eluted at 550 to 650 mM KCl and His_6-QKI at 1.1 M KCl.

Electrophoretic mobility-shift analysis

Radiolabeled RNA (100 pM) was incubated with recombinant His_6-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 hr. 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 hr at room temperature, using 0.5x TBE as running buffer.

Radiolabeled RNA (1 nM) was incubated with recombinant His_6-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 $MgCl_2$, 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 hr 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 hr before crosslinking.

miRNA profiling

miRNAs were extracted from FLAG/HA-AGO immunoprecipitates as described in Meister et al. (Meister et al., 2004). miRNAs from immunoprecipitates and the lysate were cloned and Solexa-sequenced (Hafner et al., 2008) 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 hr 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 (Andrus and Kuimelis, 2001). Briefly, in a 30 μl volume, 40 μg of purified total RNA were incubated for 16 hr 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 × 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 hr 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 500 × g 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 13,000 × g. 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 Na_2HPO_4, 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 hr 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 $MgCl_2$, 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 $MgCl_2$, 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-^{32}P-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, nonradioactive 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 (Hafner et al., 2008). The first step, 3′ adaptor 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 adaptor sets, followed by PCR to introduce primers compatible with 454 sequencing; UV 365 nm crosslinked sample RNAs were processed using Solexa sequencing adaptor sets. Depending on the amount of RNA recovered, 5′-adaptor-3′-adaptor 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.

Bioinformatic analysis

Adaptor 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., 2008). 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 adaptor to the 3′ end of each sequence read, allowing for the possibility that the adaptor was not completely sequenced (Software can be downloaded from http://www.mirz.unibas.ch/restricted/clipdata/RESULTS/index.html). 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 (Berninger et al., 2008)) were discarded because they would probably map to multiple genomic locations. The remaining sequences were mapped to the hg18 version of the human genome assembly that was downloaded from the University of California at Santa Cruz (http://genome.cse.ucsc.edu) and to a database of sequences whose function (rRNA, tRNA, sn/snoRNA, miRNA, mRNA, etc.) is already known. These were obtained from the sources specified in (Berninger et al., 2008). The Oligomap algorithm (Berninger et al., 2008) 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 (Wu and Watanabe, 2005) 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 S1B and 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 S2. 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 (Galarneau and Richard, 2005). 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 [van Nimwegen, 2007]) 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 S2). 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 S2). 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, second, 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 (Siddharthan et al., 2005). 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., 2008), section 4.1). Replicating this procedure multiple times yielded very similar results (not shown). For each protein, we selected the largest cluster of weight matrices, i.e., the cluster that contained most of the weight matrices that we obtained in replicate runs, and created the final weight matrix by summing up the counts for each nucleotide of the weight matrices belonging to this cluster. Since the clustering procedure also allows the fusion of only partially overlapping weight matrices, the resulting weight matrices are typically longer (roughly 10 nucleotides) than the motif length that we imposed in individual runs, and can contain stretches of low information content. We therefore selected for each RBP, the window with highest information content. For PUM2 and QKI, the length of this window was 8 and 6 nt, respectively, in accordance with the known or expected consensus motifs (Galarneau and Richard, 2005; Gerber et al., 2006), 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 (Valverde et al., 2008). To identify binding sites in PUM2 clusters of aligned sequence reads using the inferred weight matrix, we used the MotEvo algorithm (van Nimwegen, 2007), 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 S1A). 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 (Ule et al., 2006), 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 (the genomic regions are shown on the website: http://www.mirz.unibas.ch/restricted/clipdata/RESULTS/index.html). 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 S3, blue traces). The binding site was defined to be UGUANAUA for PUM2, ACUAA or AUUAA for QKI and CAU or two CAUs separated by no more than 10 nucleotides for the IGF2BPs. To estimate the number of sites expected by chance, we generated 1000 sets of random sequences with the same nucleotide frequencies as the CCRs (dinucleotide shuffling for PUM2 as well as QKI and mononucleotide shuffling for the IGF2BPs, due to the small length of the binding motif). For all proteins, the CCRs are clearly enriched in the respective binding motifs. The enrichment is strongest for PUM2, which has the longest recognition motif. For the IGF2BPs, the enrichment for the CAU-spacer-CAU motif is much stronger than for the CAU motif due to the clustering of the CAU motif (see previous section). For PUM2, we additionally determined the enrichment only for the first half of motif

UGUA. This short motif is clearly enriched and is contained in more than 72 percent of all CCRs, suggesting the presence of other variants of the PUM2 motif besides the consensus UGUANAUA.

Analysis of siRNA knockdown experiments

We imported the CEL files into the R software (http://www.R-project.org) using the BioConductor affy package (Gentleman et al., 2004). The transcript probe set intensities were background-corrected, adjusted for nonspecific binding and quantile normalized with the GCRMA algorithm (Wu, 2006). 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 noncrosslinked 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 toward 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 nonoverlapping 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 noncrosslinked 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 "Filtering to remove unspecific "background" clusters for AGO and TNRC6" section above) with those for which we had destabilization and AGO-IP Affymetrix microarray measurements. We then constructed, for each miRNA, three nonoverlapping 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 S7). 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 S7) that occurred more than 25 times in a given data set. Finally, we compared the destabilization and AGO-binding of crosslinked and noncrosslinked single miR-124 and miR-7 seed matches (Figure S7). 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., 2008; Landgraf et al., 2007).

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 (Berninger et al., 2008). To determine whether the enriched motifs may correspond to miRNAs, all significantly enriched motifs (with a posterior > = 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 1.0e-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 (Landgraf et al., 2007), 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 nonoverlapping 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 S6F) and second counting of the number of seed match pair occurrence in the extended clusters for 100 sets of 20 randomly selected miRNAs (Figure S6H). A visualization of seed match pair occurrence is shown in Figure S6G.

Properties of crosslinked and noncrosslinked miRNA seed matches

For the analyses whose results are presented in Figure S7 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 noncrosslinked 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 noncrosslinked 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 noncrosslinked 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., 2007).

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 (Grimson et al., 2007). 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 3′end 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 (Hofacker, 2003) 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 noncrosslinked seed matches using Wilcoxon's rank sum test.

Codon adaptation index around crosslinked and noncrosslinked seed matches

We compared the Codon Adaptation Index (CAI) (Sharp and Li, 1987) around crosslinked and noncrosslinked 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 (Su et al., 2004). 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 noncrosslinked 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 noncrosslinked counterparts, we performed local polynomial regression (Cleveland et al., 1992), fitting the distance between the seed matches and the STOP codon to the CDS length (Figure S7M, 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 noncrosslinked seed matches (Figures S7K-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 toward 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.

SUPPLEMENTAL REFERENCES

Andrus, A., and Kuimelis, R.G. (2001). Base composition analysis of nucleosides using HPLC. Current Protocols in Nucleic Acid Chemistry *Chapter 10*, Unit 10 16.

Berninger, P., Gaidatzis, D., van Nimwegen, E., and Zavolan, M. (2008). Computational analysis of small RNA cloning data. Methods *44*, 13–21.

Cleveland, W.S., Grosse, E., and Shyu, W.M. (1992). Local regression models. In Statistical Models in S, J.M. Chambers, and T.J. Hastie, eds. (Wadsworth & Brooks/Cole).

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

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

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80.

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

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol. Cell *27*, 91–105.

Hafner, M., Landgraf, P., Ludwig, J., Rice, A., Ojo, T., Lin, C., Holoch, D., Lim, C., and Tuschl, T. (2008). Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods *44*, 3–12.

Hofacker, I.L. (2003). Vienna RNA secondary structure server. Nucleic Acids Res. *31*, 3429–3431.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M., et al. (2007). A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. Cell *129*, 1401–1414.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol. Cell *15*, 185–197.

Sharp, P.M., and Li, W.H. (1987). The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. *15*, 1281–1295.

Siddharthan, R., Siggia, E.D., and van Nimwegen, E. (2005). PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny. PLoS Comp. Biol. *1*, e67.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA *101*, 6062–6067.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B.J., and Darnell, R.B. (2006). An RNA map predicting Nova-dependent splicing regulation. Nature *444*, 580–586.

Valverde, R., Edwards, L., and Regan, L. (2008). Structure and function of KH domains. FEBS J. *275*, 2712–2726.

van Nimwegen, E. (2007). Finding regulatory elements and regulatory motifs: a general probabilistic framework. BMC Bioinformatics *8*, S4.

Wu, Z., Irizarry, R.A., Gentleman, R., Murillo, F.M., and Spencer, F. (2004). A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Johns Hopkins University, Dept. of Biostatistics Working Papers *Working Papers*, Working Paper 1.

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

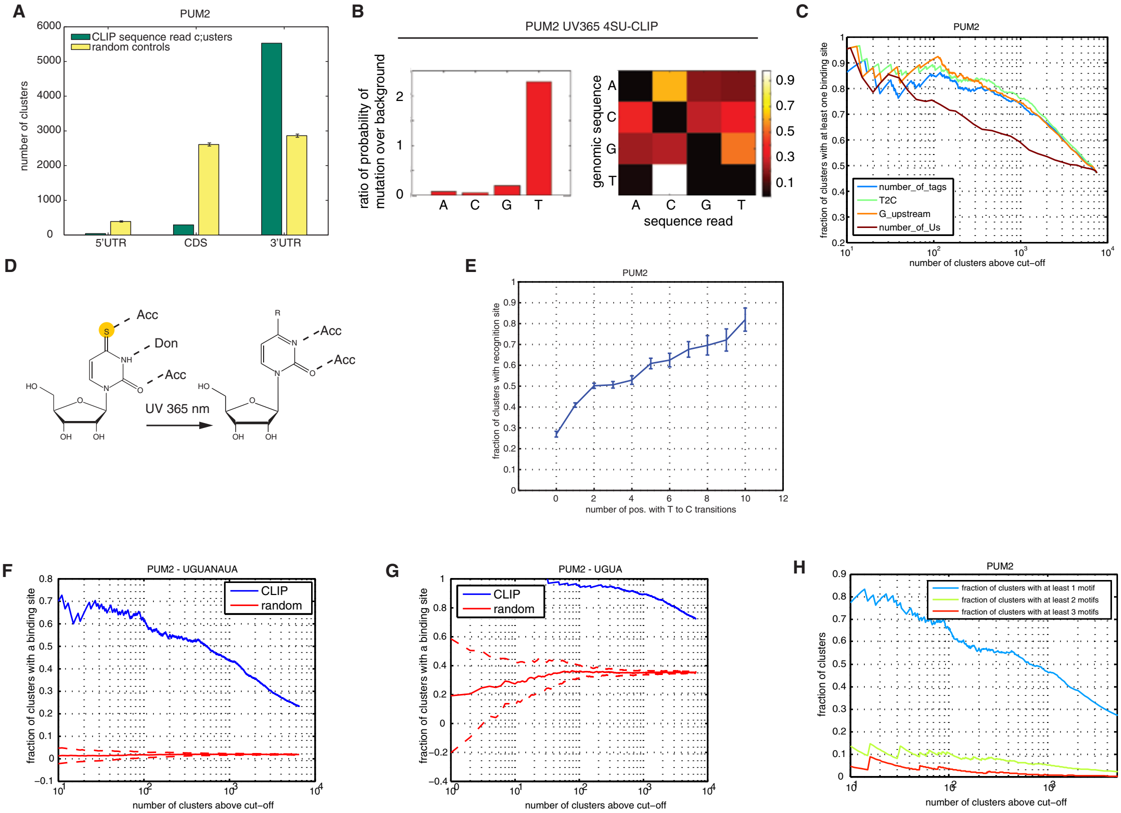


Figure S1. Analysis of PUM2-PAR-CLIP clusters, Related to Figure 2

(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 toward each of the three others. In the right panels, white indicates high mutation frequency toward a particular nucleotide. 4SU-PAR-CLIP yields about a 15-fold increased mutation preference for T, nearly always to C.

(C) Fraction of clusters containing the PUM2-recognition motif, versus the total number of clusters above a given cut-off on a particular property as indicated in each figure legend (G upstream: number of sequence reads with a G at position −1; T to C: number of sequence reads with a T to C mutation; number of sequences: total number of sequence sequence reads in the cluster, number_of_Us: number of uridines in the sequence read cluster). For each cut-off on a given property, the fraction of clusters with at least one binding site above the given cut-off is shown. Cut-off increases from right to left. The best signal is obtained by sorting according to the frequency of crosslinking events.

(D) The increase in T to C transitions after 4SU-protein crosslinking can be rationalized by structural changes in donor/acceptor properties of 4SU after crosslinking to proximal amino acid side chains and subsequent incorporation of dG rather than dA in the reverse transcription; R representing a side chain.

(E) Fraction of clusters with the recognition element (as indicated) for PUM2 versus the number of distinct crosslinking sites within a cluster indicated by a T to C change. The fraction of sites containing at least one recognition motif rises with the number of crosslinking sites.

(F–H) Enrichment of binding motifs for PUM2 for the consensus motif UGUANAUA (F) as well as the short variant UGUA (G) compared to CCRs with randomized sequences. Panel (H) shows the fraction of clusters with at least one, two or three UGUANAUA motifs. Most clusters contain only one binding site.

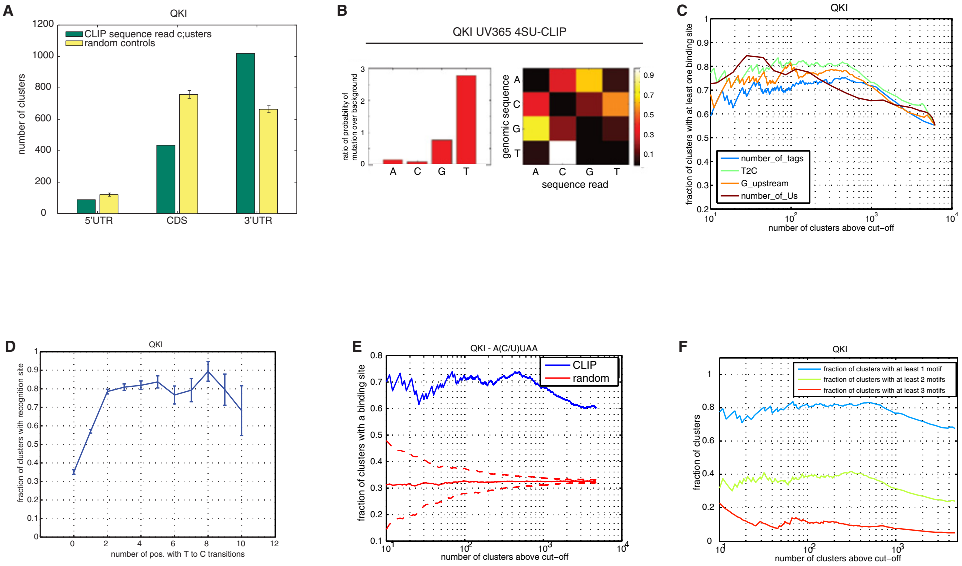


Figure S2. Analysis of QKI-PAR-CLIP clusters, Related to Figure 3

(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 toward each of the three others. In the right panels, white indicates high mutation frequency toward 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.

(F) Panel shows the fraction of clusters with at least one, two or three motifs. A significant fraction of clusters contains two or more binding sites.

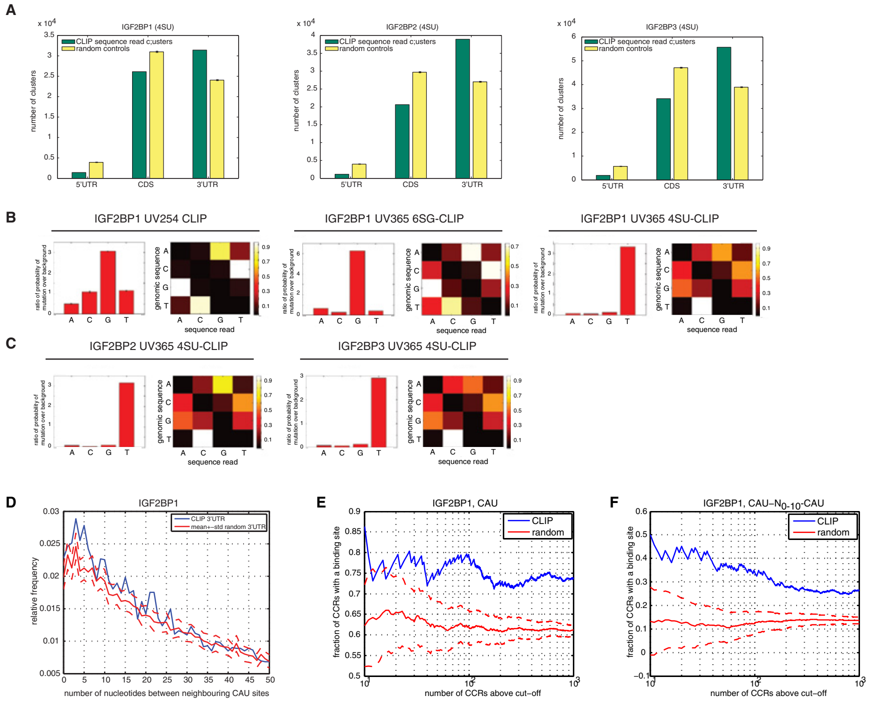


Figure S3. Analysis of IGF2BP1-3-PAR-CLIP clusters, Related to Figure 4

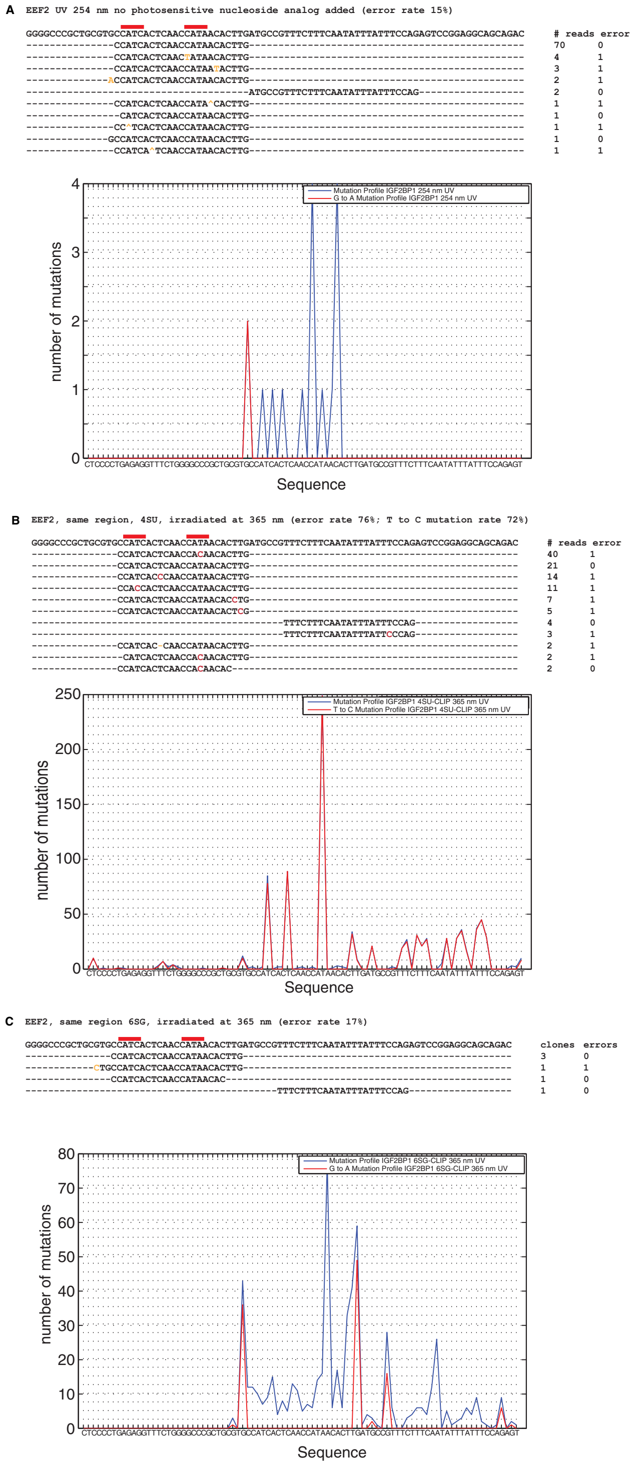
(A) Analysis of the transcript regional preferences and the mutational pattern of crosslinked sequences of IGF2BP1-3. The number of exonic sequence read clusters annotated as derived from the 5′UTR, CDS or 3′UTR of a target transcript is shown (green bars). Yellow bars show the expected location distribution 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 toward each of the three others. In the right panels, white indicates high mutation frequency toward 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.

(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
```

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
```

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
```

Figure S4. 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).

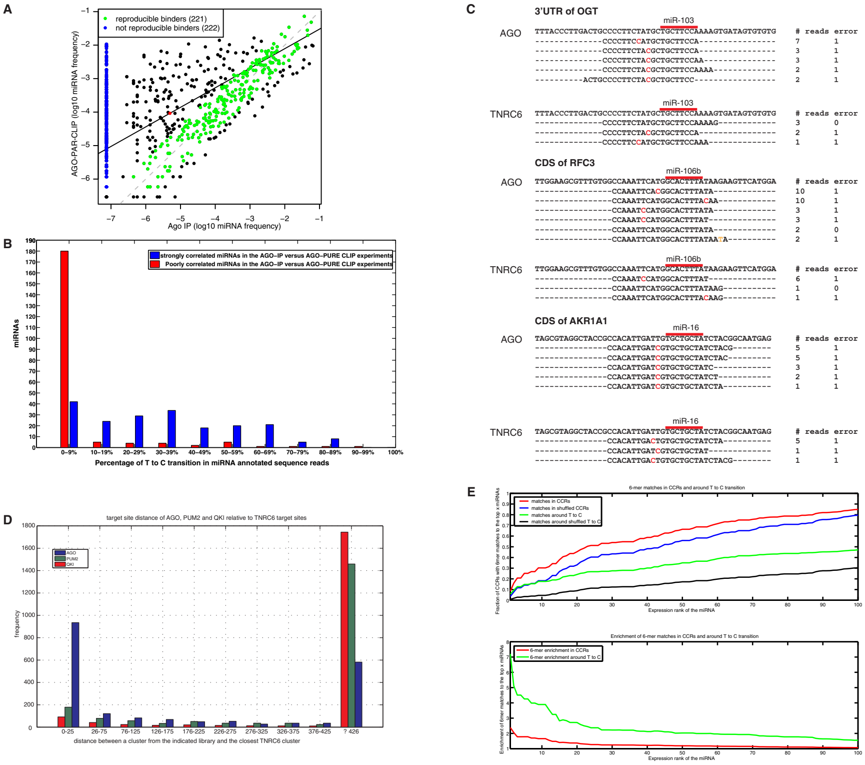


Figure S5. AGO-protein family PAR-CLIP, Related to Figure 5

(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 versus 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 versus 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).

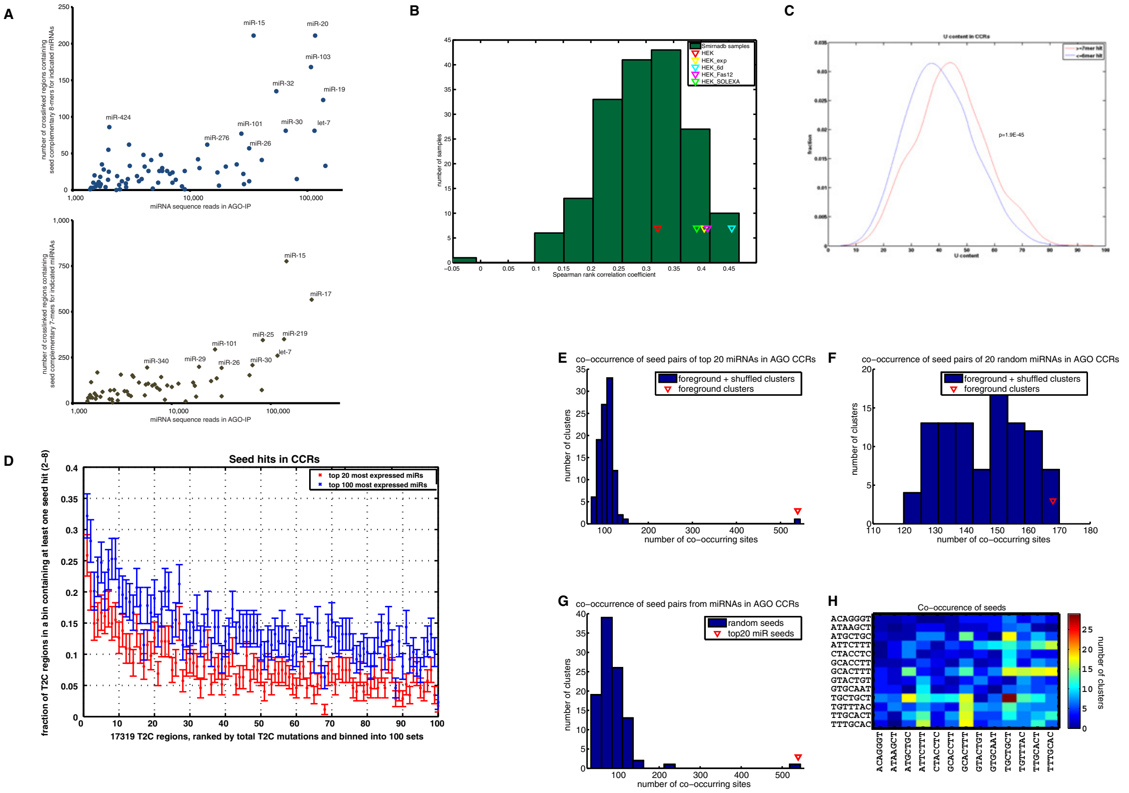


Figure S6. Seed complementary sequences from abundant HEK293 miRNAs are enriched in AGO-PAR-CLIP CCRs. Related to Figure 6

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 = 1.9E-45).

(D) The number of crosslinking events correlates with the enrichment of the CCRs in the putative binding sites for the most abundantly expressed miRNAs. The frequency of the most strongly enriched miRNA seed motif (complementary to positions 2-8 of the miRNAs) was determined in the 17,319 AGO CCRs, which were sorted by the number of U-to-C changes and grouped into bins of 100. The frequency of miRNA seed-complementary motifs in the CCRs decreases with the number of U-to-C mutations in the clusters corresponding to these CCRs.

(E) Number of pairs of nonoverlapping seed (pos. 2-8) matches for the 20 most abundantly expressed miRNAs in HEK293 cells in the crosslinked regions (red triangle) and in control regions (100 sets of dinucleotide shuffled crosslinked regions). Only the experimental set shows enrichment of miRNA pairs.

(F) Number of co-occurring pairs of miRNA seed matches in the AGO crosslinked regions and the shuffled control regions for 20 randomly chosen miRNAs.

(G) Number of co-occurring pairs of miRNA seed matches in the AGO crosslinked regions for 100 sets of 20 randomly chosen miRNAs.

(H) Heat map representation of miRNA seed match co-occurrence. Only miRNA seed matches were counted that did not overlap and could therefore be bound simultaneously by two AGO-proteins. The scale indicates the absolute number of co-occurring pairs. Matches to the seed of miR-17 co-occur with matches to the seed of miR-19/miR-130/miR-301/miR-30/miR-15/miR-16. miR-16 seed matches have the tendency to co-occur with themselves.

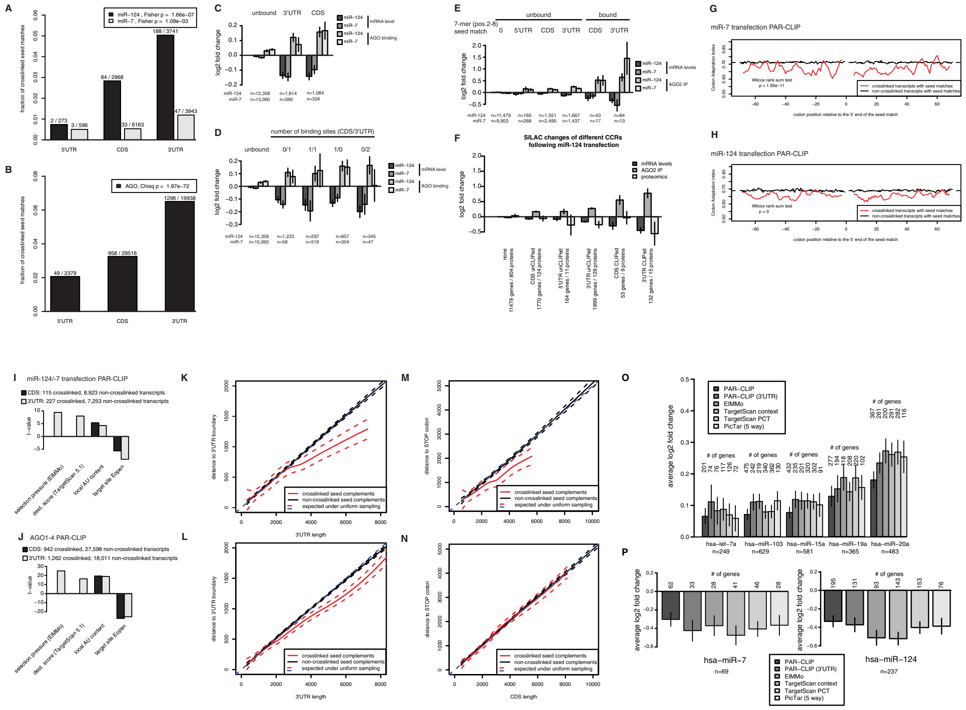


Figure S7. Properties of CCRs containing miRNA seed complementary sites, Related to Figure 7

(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 gray 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 gray bars) was determined by comparing transcripts enriched by AGO2-IPs of mock transfected and miR-124 and miR-7 transfected HEK293 cells (Hausser et al., 2009). 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 (Baek et al., 2008)) 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 noncrosslinked transcripts, respectively. (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 noncrosslinked 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 noncrosslinked 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 noncrosslinked 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 S, 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.